DOCUMENT RESUME

ED 314 236 SE 050 680

AUTHOR Raizen, Senta A.; And Others
TITLE Assessment in Elementary School Science Education.
INSTITUTION Biological Sciences Curriculum Study, Boulder, Colo.; National Center for Improving Science Education, Washington, DC.; NETWORK, Inc., Andover, MA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 89
GRANT RI88B80001
NOTE 164p.; For related reports, see SE 050 679-681.
PUB TYPE Viewpoints (120) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC07 Plus Postage.
DESCRIPTORS Achievement; Attitude Measures; *Elementary School Science; Evaluation; *Evaluation Methods; *Evaluation Problems; *Evaluation Utilization; Objectives; *Program Evaluation; Science Education; Science Programs; *Science Tests; Student Evaluation

ABSTRACT

This report discusses the purposes and nature of various forms of assessment that can be used to enhance, support, and monitor the progress of science learning in elementary school classrooms. Chapters included are: (1) "Introduction" (describing assessment priorities); (2) "Issues in Assessment" (discussing validity, correspondence between curriculum, instruction, and test, use of results, and program assessment); (3) "Assessment of Student Learning" (dealing with the contents, methods, usages, and the attitude assessments); (4) "Assessment of Program Features" (discussing the reason, indicators, usages, targets, and self-assessment) and (5) "Improving Assessments in Elementary Science Education" (including improvement goals, starting point, and systematic approach). Eight fundamental organizing concepts for elementary school science are described and exemplified in the appendix. (YP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314236

The National Center for Improving Science Education

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Senta Raizen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Assessment in Elementary School Science Education

SE 050 680

A PARTNERSHIP OF THE NETWORK, INC. AND THE BIOLOGICAL SCIENCES CURRICULUM STUDY (BSCS)

The NETWORK, Inc.

BSCS
The Natural Selection

ASSESSMENT IN

ELEMENTARY SCHOOL SCIENCE EDUCATION

by

Senta A. Raizen
Joan B. Baron
Audrey B. Champagne
Edward Haertel
Ina V.S. Mullis
Jeannie Oakes

1989

The National Center for Improving Science Education

A Partnership of

The NETWORK, Inc.
Andover, Massachusetts
Washington, DC

and

The Biological Sciences Curriculum Study
Colorado Springs, Colorado

This report is based on work sponsored by the Office of Educational Research and Improvement, U.S. Department of Education, under Grant number R168B80001. The content of this report does not necessarily reflect the views of OERI, the Department, or any other agency in the U.S. Government.

THE NATIONAL CENTER FOR IMPROVING SCIENCE EDUCATION
ADVISORY BOARD MEMBERS

William Baker
Chairman (Retired)
AT&T Bell Laboratories
Murray Hill, NJ

Richard M. Berry
National Science Foundation
(Retired)
Bowie, MD

David P. Crandall
President
The NETWORK, Inc.
Andover, MA

Sally Crissman
Science Teacher and Lower
School Head
Shady Hill School
Cambridge, MA

Richard Duncan
President
Association of Presidential
Awardees in Science Teaching
Whitford Intermediate School
Beaverton, OR

David Kennedy
President
Council of State Science
Supervisors
Olympia, WA

Douglas Lapp
Executive Director
National Science Resources Center
National Academy of Science
Washington, DC

Jerome Pine
Professor of Biophysics
California Institute of Technology
Pasadena, CA

Andrew Porter
Director
Wisconsin Center for Educational
Research
University of Wisconsin
Madison, WI

Mary Budd Rowe
Professor
College of Education
University of Florida
Gainesville, FL

Kenneth Russell Roy
National Director
National Science Supervisors
Association
Glastonbury Public Schools
Glastonbury, CT

F. James Rutherford
Chief Education Officer
American Association for the
Advancement of Science
Washington, DC

Thomas P. Sachse
Manager
Mathematics, Science and
Environmental Education
California Department of Education
Sacramento, CA

Mare Taagepera
Director
Science Education Program
University of California
Irvine, CA

Laurence Vickery
General Director for Regional
Personnel Administration
General Motors Corporation
Warren, MI

ASSESSMENT PANEL

Joan Boykoff Baron
Director
Connecticut Assessment of Educational Progress Program
Connecticut State Department of Education
Hartford, CT

Audrey B. Champagne
Senior Program Director
Office of Science and Technology Education
American Association for the Advancement of Science
Washington, DC

Edward Haertel
Associate Professor
School of Education
Stanford University
Stanford, CA

Ina V.S. Mullis
Associate Director
National Assessment of Educational Progress
Princeton, NJ

Jeannie Oakes
Social Scientist
The Rand Corporation
Santa Monica, CA

Senta A. Raizen, Chair
Director
The National Center for Improving Science Education
Washington, DC

CONTENTS

FOREWORD i

CHAPTER I. INTRODUCTION 1
A Science Classroom 1
Why Worry About Assessment? 3
Assessment Priorities 10

CHAPTER II. ISSUES IN ASSESSMENT 15
Testing What Matters 15
Correspondence Between Curriculum, Assessment for Instruction, and Assessment for Monitoring 29
The Uses of Assessment 35
Assessing Science Programs 44

CHAPTER III. ASSESSMENT OF STUDENT LEARNING 47
What to Assess 47
How to Assess 54
Using Assessments in Elementary Science Education 69
Assessing Attitudes and Dispositions 75

CHAPTER IV. ASSESSMENT OF PROGRAM FEATURES 79
Why Assess Elementary School Science Programs? 79
Balancing the Effects of Assessment on Science Programs 79
Enhancing the Policy Relevance of Science Assessments 82
What Program Characteristics Should Be Assessed? 83
Effective Self-Assessment of the Science Program 91

CHAPTER V. IMPROVING ASSESSMENTS IN ELEMENTARY SCIENCE EDUCATION 95
Improvement Goals 95
A Starting Point 99
A Systemic Approach 100

REFERENCES 123

APPENDIX

FOREWORD

This report is one of a series produced by the National Center for Improving Science Education. The Center's mission is to promote changes in state and local policies and practices in the science curriculum, science teaching, and the assessment of student learning in science. To do so, the Center synthesizes and translates the findings, recommendations, and perspectives embodied in recent and forthcoming studies and reports in order to develop practical resources for policymakers and practitioners. Bridging the gap between research, practice, and policy, the Center's work is intended to promote cooperation and collaboration among organizations, institutions, and individuals committed to the improvement of science education.

The synthesis and recommendations on assessment in this report were formulated with the help of the study panel whose members are listed in the front (page iii) of this report. We gratefully acknowledge the help given to us by many individuals who have supplied materials and made recommendations and suggestions for the text of the report. While the list would be too long to acknowledge individually, we wish to give special thanks to Richard Berry, formerly of the National Science Foundation, and Elizabeth Badger, of the Massachusetts Department of Education, for their contributions to the text of this report. We also thank Richard Shavelson and the other reviewers of the report for their critical comments which helped to improve it. Thanks are also due to the support of the Center's monitors at the U.S. Department of Education, John Taylor and Wanda Chambers.

Two other panels have produced companion reports on curriculum and instruction and on teachers and teaching. A summary report integrating all three of these documents will be prepared and will be available from the Center. This integrative report will be supplemented by implementation guides for state and district policymakers and practitioners, and by guidelines especially tailored for additional audiences including teachers, principals, school boards, parents, and teacher educators.

The Center, a partnership between The NETWORK, Inc. of Andover, Massachusetts and the Biological Sciences Curriculum Study (BSCS) of Colorado Springs, is funded by the U.S. Department of Education's Office of Educational Research and Improvement. Members of its Advisory Board are listed on page iii of this report. For copies of this report or further information on the Center's work, please contact Senta Raizen, Director, National Center for Improving Science Education, 1920 L Street, NW, Suite 202, Washington, DC 20036, or Susan Loucks-Horsley, Associate Director, National Center for Improving Science Education, The NETWORK, Inc., 290 South Main Street, Andover, MA 01810.

I. INTRODUCTION

A Science Classroom

"How do seeds live? Can seeds grow way, way deep in the ocean and make seaweed?" "How do seeds get inside of watermelons?" "Hey! How do they make watermelons without seeds in them?" "How do seeds grow plants?" These were some of the many questions asked by Ms. Lopez's second graders. Today, they are thinking about seeds, the topic they are about to study, and Ms. Lopez is keeping track of these questions on a chart titled: "Questions We Have about Seeds." Another chart titled: "What We Know About Seeds" is filled with such statements as: "Seeds grow in gardens," "You can eat sunflower seeds," and "Carrots don't have seeds." These charts are referred to time and again as the teacher encourages questions to develop concepts and change opinions. Ms. Lopez uses the children's questions and comments to decide that the children are ready for a "seed walk."

The next morning, the students go to a nearby field to collect seeds. Each child, besides carrying a collection bag, wears an adult sock over one shoe and pulled up to the knee, providing a fuzzy surface to which seeds can cling. When the children return from the walk, they each select one seed to study carefully with a hand lens. After each child makes observations about what the seed looks, feels, and smells like, and guesses how it might have traveled, the child makes a presentation to the group in the meeting circle. The teacher keeps track of the kinds of seeds discussed by taping the specimens onto a chart. After the children tally the number of the different kinds of seeds the group has collected, they develop picture graphs of the results.

That evening, after the "seed walk," Ms. Lopez reflects on the differences in the children's understandings of the structure and function of seeds. She notes which children easily made observations and which ones had more difficulty, which children made more obvious or more unexpected responses, and which children seemed comfortable using the lens for examining their seeds and which ones seemed more awkward. As she thinks of the multiple activities for the next day, Ms. Lopez uses her notes to place children in groups so that their discussions will prompt and challenge one another's inquiry.

The next day, some groups count the seeds on their socks and then plant them in large plastic baggies, watering and setting them in the window area. In the days that follow, they will be encouraged to observe the germination process carefully and compare the total number of seeds with the number that sprouted by making "ratio" sentences. Ms. Lopez invites other children to compare sizes of seeds by outlining the seeds on graph paper and then counting the number of graph squares each seed covers. The students discover there is a great diversity of sizes and shapes in different kinds of seeds, and that the same kind of seed has variations in size and shape.

Still other groups choose to continue working in the "seed journals" that she requires all to keep. They are either to paste in or draw the specimen and then "write" about three seeds of their choice, including the same sorts of observations they shared earlier in meeting circles. Since students of this age have a range of "sentence" writing capabilities, the teacher meets with each child to discuss that individual's observations and writing. She uses the journals and evidence from the meetings to monitor the level of understanding the children have of such concepts as diversity and cycles.

Ms. Lopez's class spends most of the week working on this science topic, incorporating writing and math, as well as inquiry-based science activities. Other activities she will do with the children include: a fiction story about how a native American girl uses seeds and plants, a garden song, and drawing the seedlings as they sprout. Her thematic active learning approach is similar to that she observed and practiced during a year of induction, when she was coached by a mentor as she tried her first interdisciplinary unit.

In successive lessons, Ms. Lopez will call groups together and, based on their explorations, ask several questions. As she records the responses, Ms. Lopez will ask the children to clarify their answers. Eventually, she will introduce some new vocabulary information that helps the students to reflect on their developing concepts. Some of the children may not be sure about the new information; they will need more time to talk about it and do some additional testing of their ideas to help make the new information part of their personal understanding of seeds. Last year when she did this unit, for example, several youngsters insisted that the lima bean embryos they discovered earlier would grow into lima bean plants even without the "seed halves" attached. They were convinced that the embryos could "eat" the soil and water and grow into an "adult" lima bean plant. Through careful questioning, Ms. Lopez was able to guide these children to design a test of their beliefs. She found that these children changed their point of view after they conducted the investigation, and that they now had some additional questions to pursue.

After several weeks of studying seeds, Ms. Lopez recognizes that the children have learned a great deal about such science concepts as diversity, life cycles, and structure and function. They have become adept "observers" and ask questions of each other and of Ms. Lopez concerning these developing concepts. Ms. Lopez knows they will soon be ready to apply these new levels of knowledge and skills to other science areas. The children will, as a group, construct a booklet on how to plant seeds and care for the seedlings. Ms. Lopez will keep notes on the progress of individual children and the class as a whole. This will then help her plan and design more effective science instruction to use in future classes. It will also provide the source material that will enable her to make more formal assessments in report cards, in conferences with parents, and -- for the class as a whole -- to Mr. Sandowski, the 3rd grade teacher.

Why Worry About Assessment?

Should Ms. Lopez be concerned about how she assesses her students' progress? About the district or state science test that may be mandated for her students next year? It seems obvious that any effort to help elementary schools do a better job in science education must concern itself with improving curriculum and instruction and with the quality of teaching and the competence of teachers in science. But why worry about assessment? There are three important reasons:

1. Assessment can be a helpful tool for the teacher to guide instruction and make it more effective.
2. Assessment can impress on students, school staff, and parents the importance of science learning.
3. Assessment can be used as a policy tool to monitor the outcomes of science instruction and help improve science programs.

Confusion often prevails over these different purposes of assessment, particularly the distinction between assessment for instructional purposes and assessment for monitoring purposes. Before discussing these distinctions, however, we note a fourth reason that assessment deserves a high place on any improvement agenda:

4. Assessment can exert a powerful influence on curriculum and instruction, for good or ill. As mandates for assessment grow, it becomes critical to establish correspondence between the goals of science education, the curriculum, and the tests and other means of assessment used to establish what children have learned and can do in science. Assessments must support Ms. Lopez's teaching, not undermine it. This is true no matter what the assessment purpose.

The following discussion elaborates somewhat further on each of these four reasons for giving attention to assessment.

Assessment in the Service of Instruction

Teachers may use tests or other forms of assessment for a variety of instructional purposes. Ideally, as in the case of Ms. Lopez, these include: (1) finding out what information and constructs students bring to a science lesson so as to build on their prior knowledge and conceptions; (2) establishing, after some sustained period of instruction, what students have learned in order to shape subsequent teaching; (3) placing students in productive learning groups to make instruction more effective; (4) motivating students to learn assigned material; (5) communicating to students the teacher's expectation of what they are to learn; and (6) documenting what students have learned in order to inform them, parents, and subsequent teachers of individual and group progress. Thus, at its best, assessment can be a powerful tool for focusing instruction and providing valuable information about how to increase learning. If it is incorporated into instruction in thoughtful ways, assessment can provide teachers with the feedback they need to help their students.

Unfortunately, too few teachers use assessment as Ms. Lopez does. The most common use of assessment in the classroom is to assign grades to individual students. Assessments to support instruction are seldom done. Moreover, to serve the narrow grading purpose, teachers generally develop their own tests or rely on tests embedded in the students' textbooks and accompanying teaching materials. In either case, there is reason for concern about the quality of these tests given the lack of pertinent background and training in assessment issues and techniques that would allow teachers to evaluate and construct tests (Dorr-Bremme and Herman, 1986). Understandably, teachers seldom choose to use standardized or mandated tests (or the results from these tests) for their own purposes. Quite rightly, they see these tests as largely irrelevant, except for rough student placement at the beginning of the school year.

Assessment as a Conveyor of Expectations

Expectations about science learning in the elementary grades operate at two levels: expectations by the public -- including parents, school boards, and policymakers -- about the importance of science education in the early years of formal education and expectations of classroom teachers about their own students. It has been argued that science, despite the many recommendations urging that it become a new "basic" (e.g., National Commission on Excellence in Education, 1983; National Science Board, 1983), will not assume any significant importance in the elementary-school curriculum until there is the same kind of stress on testing science knowledge as there is on testing reading and arithmetic skills. In fact, more and more states, as they reform their elementary science programs, are mandating science assessments in 4th (sometimes 3rd or 5th) grade (Blank and Espenshade, 1988). Although it might be regrettable, given the currently limited ability to assess important learning outcomes in science, it appears that the importance of this subject, as of other subjects, is gauged by the extent of student testing that takes place.

The age-old student query: "Will it be on the test?" demonstrates the power of assessment to convey teacher expectations of what is to be learned. This makes the current limitations of science testing, discussed in greater detail in subsequent chapters, doubly troubling, since these limitations act not only on teachers in narrowing what they choose to emphasize but also on students as they attempt to concentrate their study on what is most likely to pay off in terms of high test scores and good grades.

Assessment as a Policy Tool

As assessment moves beyond Ms. Lopez's classroom, its purposes include: (1) providing an indicator of the condition of education -- whether in the nation, a state, a district, or a school -- through periodic monitoring of student learning; (2) accounting for monetary and human investments made in education through assessing the results achieved in student learning; and (3) evaluating the effectiveness of particular programs with respect

to learning outcomes for students.

As for to monitoring educational outcomes, policymakers and the public they represent generally are interested in answers to broad questions: What is the general level of student accomplishment, and what are the percentages of students who attain different levels of achievement? Are students today doing as well, say, in science as students did a decade or two ago? Are different population groups showing different achievement levels for example, U.S. students compared to students in other industrialized countries, or students in different regions or states of the country or in different school districts within a state, or students from different ethnic or socioeconomic backgrounds? At times, these questions may take on a normative character; witness the current concerns with the perceived mismatch between what students are learning in school and the needs of tomorrow's labor market for thoughtful, creative individuals who can solve problems under changing conditions and in new contexts (Committee on Science, Engineering, and Public Policy, 1984; Scheuer, 1987; Twentieth Century Fund, 1983).

The impetus for assessing student outcomes for indicator and accountability purposes generally comes from administrators within the district or state or from policymakers at various levels of government (local, state, national) who are also the main audience, although media interest may become quite high. Since this type of assessment is generally externally mandated, it tends to use externally constructed, standardized tests. These may be commercially available tests or tests specially developed by a state or district. Although the tests are often administered to all students in selected grades in a given district or state, this is not necessary if a large enough representative sample can be drawn to allow generalization to the whole population. There are two general problems: First, administrators and policymakers need to ascertain that a test actually measures the student outcomes of interest to them. Second, care needs to be taken with the reference standards used to interpret test results. Referencing test scores against national norms may tell very little about the quality of student science learning and the development of science understanding in a school, district, or state.

Policymakers at every level also are interested in bringing about improvement, particularly if the information on educational outcomes proves disappointing (Oakes, 1986; Womer, 1981). From this perspective, policymakers can be likened to the chief executive officers of major businesses who conduct product evaluations to guide decisions about resource allocations. Assessment results can point to designs for effective programs, improved approaches for retraining teachers, or the effectiveness of magnet schools. Assessment results can be used by school administrators to make decisions about the success of special curriculum approaches, particular instructional strategies, and teacher selection procedures. Of course, this necessitates more than assessing the outcomes of education. A theory is necessary that makes causal connections between educational resources and processes and student outcomes, and the most critical resource and process factors posited by the theory must then be assessed to provide guidance about which of these need to be changed to achieve improved outcomes.

The component of assessment necessary for making improvements, the evaluation of programs, usually is of interest to administrators immediately concerned with the effectiveness of alternatives available in, say, science education -- a smaller audience than that interested in information on student outcomes. Assessments of program characteristics often are designed and conducted by university researchers and curriculum developers. If student learning is specified as a criterion of program effectiveness, these assessments should use tests that address the specific goals and content of the program. Judgments whether the goals and content themselves are of high quality (i.e., embody what students ought to learn) can and probably should be made independently, but it is precisely in these judgments and in the fit between goals, curriculum, and assessment that severe problems arise.

Correspondence Between the Goals of Science Education, the Science Curriculum, and the Assessment of Student Learning

---

Whether assessment of student learning in science is used to inform instruction in the classroom or to formulate broader policy at the district, state, or national level, it is critical that the domains probed by any assessment that purports to be comprehensive range across all the important educational goals. Items and exercises constituting such a test need to assess three component areas of science learning: (1) factual and conceptual knowledge; (2) skills in the use of apparatus and equipment necessary to do science, including hands-on performance and the science thinking skills and general thinking skills used in reasoning and problem solving in science; and (3) the disposition to apply science knowledge and science-based skills outside the classroom.

Of all the desired outcomes, the acquisition of factual knowledge is easiest to assess through the usual multiple-choice items that posit one "correct" response. Adequate assessment of the other competencies may require observation, open-ended responses, allowing multiple answers, and following student progress over time. Not surprisingly, therefore, tests used to determine what students have learned tend to be dominated by items eliciting factual recall, whereas assessment of science understanding, science skills, and the disposition to apply science knowledge and skills tend to be neglected. If the outcomes of tests emphasizing factual knowledge are used to make changes in instruction or curriculum, the changes are likely to be in the direction of narrowing science education in favor of learning facts, with a concomitant deemphasis on such goals as learning how to think about questions in science and how to carry out activities that address science questions. As assessment of science learning in elementary school becomes more widely instituted -- possibly in an effort to establish science as a basic -- this problem must receive concentrated attention.

If assessment results are to reflect what students have learned as a result of their science instruction, a second requirement is that the assessment must be matched to the specific curriculum planned for a given setting or, if it can be determined, the curriculum

actually delivered to the students. No such match is necessary if the intent of the assessment is to monitor the general state of student knowledge and competence in science, as in past assessments conducted by NAEP (National Assessment of Educational Progress) and IEA (International Association for the Evaluation of Educational Achievement). In these cases, decisions need to be made on the standards of knowledge and performance to be expected from students at a given level, regardless of curriculum, or on the core that is common to most curricula and likely to be taught to most students -- an approach exemplified by many commercially available, standardized tests.

Tests used for monitoring and accountability, because of their widespread use, are more available for review than tests used by teachers for classroom purposes. Since they usually involve large numbers of students, tests used for monitoring are designed for reliability of results; ease of administration, scoring, and analysis; and appropriate psychometric properties. Inevitable, multiple-choice (or other short-answer) items make up by far the largest fraction of these tests. As noted, tests with such characteristics lend themselves best to assessment of factual knowledge and certain circumscribed reasoning and problem-solving skills; multiple-choice tests are not suited to probing achievement and performance that involve generative thinking and open-ended responses (Anderson, 1985; Frederiksen, 1984; Ward et al., 1980). Moreover, not only are the tests limited in the types of knowledge and skills they assess; they often fail to correspond well even to that part of the curriculum they do address. They tend to assess students' general knowledge in science, and at a relatively low level at that, rather than what students have learned during some period of instruction. It should not come as a surprise, then, that some commonly used tests show little progress in the learning of science as students move through the grades. Problems with these tests are aggravated when norm-referenced standards are used to interpret test results since these norms are established to rank order individual students rather than to provide insight into the development of each student's science learning.

It is more difficult to make judgments about the quality of the tests that teachers give for instructional purposes. Presumably, if the tests are curriculum-embedded or teacher-constructed, they should match the curriculum better than do standardized tests designed explicitly to be valid across many curricula. In fact, there is little evidence on the quality of tests that teachers give within their classrooms; teacher-controlled tests may do no better at probing science knowledge, skills, performance, and dispositions that are difficult to assess through conventional testing techniques. Assessment exercises using alternative techniques able to provide insight on important but generally untested science learning require time and creativity to develop, time to administer, and training in interpretation and grading. Such exercises are not readily available to teachers. Do the schedules of most science teachers permit the investment of time and energy needed to construct their own assessment exercises? How good are these? Can elementary school teachers, many of whom -- unlike Ms. Lopez -- do not feel confident of their ability to teach science, be expected to construct tests that would inform their instruction adequately?

Assessment Priorities

The preceding discussion has focused on the various reasons for assessing student learning and competencies in science. Figure 1 summarizes these reasons:

FIGURE 1
Reasons for Assessing Student Science Learning

	Improvement	Conveying Expectations	Monitoring Status	Accountability
Individual (Teacher)	Classroom Instruction	To Students and Parents	Of Individuals and Class	
Group (Policy Makers)	Science Program in District and State	To Teachers and Administrators	Of District, State, and Nation	Effective Use of Resources

↑ ↑
Assessing Program Features

The next two sections explain why our panel has chosen to concentrate much of its effort on assessments carried out by the classroom teacher and what additional assessment issues need attention to support instructional improvement. These priorities are indicated by the shaded boxes in Figure 1.

First Priority: Assessment in the Service of Classroom Instruction

In recent years, work on assessing the quality of science education in this country has concentrated on monitoring student learning and program quality for broad policy purposes (Murnane and Raizen, 1988; National Science Board, 1987; Oakes, 1986; Raizen and Murnane, 1985; Shavelson et al., 1987). Because our Center's primary mission is to help schools and teachers improve science education at the classroom level, this report emphasizes the use of assessment to guide subsequent instruction. This emphasis implies that:

- The individual classroom, as set in the school context, should constitute the basic unit for achieving improvement in science education.
- Attention should be focused on improving the kinds of curriculum-embedded and teacher-constructed tests most often used in the classroom.
- Alternatives to traditional testing need to be an explicit part of assessing student achievement and progress in science.

Within the emphasis on the classroom and school level, we concentrate on providing teachers with better means for finding out what students have learned and can do in science. Assessments of curricular quality, teacher competence, and quality of the science program as a whole receive less emphasis since these serve as constraints or incentives at the individual classroom level rather than being under the teacher's control.

Some -- though not all -- approaches useful for improving what teachers do to assess what their students have learned are also useful for broader assessments since the

problems of probing all important domains of science education appropriate at a given age or grade level are similar, what ever the purpose of the assessment. Fortunately, the teacher has available strategies for use in the classroom that are difficult or costly to replicate with large numbers of students.

> *Ms. Lopez, for example, is able to monitor her students' progress on an ongoing basis. She uses students' individual journals, the class chart on "What We Know About Seeds," and her daily notes on the oral participation of individual students to record whether and to what extent the students are developing appropriate notions about such key principles as diversity, organization, change, and systems. She observes progress in their use of the hand lens, scales, and volume measures and in their proposals about how to test various ideas on how seeds develop into plants. As work on the plant unit goes forward, she accumulates a record of each of the children's participation in the class science activities and also of their individual oral and written work.*

Though our focus in this report is on assessing student learning, we recognize that important questions may arise for principals and district administrators on the quality and suitability of the science program within a grade or a school. Chapter IV of our report briefly addresses assessment of program characteristics. Here, too, the elements to be evaluated are analogous to those relevant for broader policy levels -- curriculum, instruction, preparation and competence of the teachers, availability and use of resources -- but the specifics of how program evaluation might be carried out at these different levels varies considerably.

Second Priority: Assessment in the Service of Policy

Assessment as a policy tool, although it has already received considerable attention in other contexts, is discussed here because of its obvious ties to improving what happens in the classroom. Although policy (and assessment conducted for policy purposes) cannot in itself cause improvement in the classroom, it can impede or facilitate improvement. (For example, district policy may make it difficult for Ms. Lopez to take her students on a field trip to the city park for the seed walk.) Moreover, as noted, some of the assessment problems are similar, whatever the level of the assessment.

However, as in the discussion of assessment at the classroom level, the major emphasis in this report is on assessing student learning, even when assessments are carried out for purposes of monitoring, accountability, and formulating state or national policy. There are three reasons for this:

- Student learning is the end purpose of education; hence, for purposes of monitoring and accountability, its assessment should take precedence over other forms of assessment.
- Understanding what students have learned and can do, the most important outcome of education, presents issues and problems that are quite distinct from assessing the resources and processes that make up program quality.
- Assessment of student learning in science is in itself a sufficiently complex and troubled area without taking up in detail the problems associated with assessing the quality of science curricula, instruction, available resources, and other critical elements of a school's science program.

Monitoring of student learning in itself, however, cannot set clear policy directions though it can point to likely options. Interpretation and discussion of results are necessary, set against what is known about policies and practices that inhibit or facilitate student learning. For example, if science is not taught or taught for only a few minutes a day, students cannot be expected to learn much science in school. Hence, knowledge about salient program features also must receive attention if effective improvements are to be made in science education. This is as so at the classroom and school levels as it is at the district, state, and national levels.

II. ISSUES IN ASSESSMENT

This chapter takes up four critical issues in assessing student learning in science. The first section discusses the unfortunate circumstance, not unique to science education but creating particular difficulties in it that the learning and competencies valued most and deemed the most important are the most difficult to assess. The second section reviews the general educational context in which assessments of science learning take place: the nature of science education; what it is and what it ought to be; and to what extent there is correspondence between the goals of science education, the science curriculum, and the assessment of what students have learned and are able to do in science. The third issue concerns appropriate and inappropriate uses of assessment inside the classroom and for policy purposes. Lastly, we argue the importance of assessing schooling factors that play a critical role in students' science learning.

Testing What Matters

Valued Outcomes of Science Education

There is broad consensus that scientific and technological literacy for all citizens stands high on the list of educational needs for the year 2000 and beyond (National Commission on Excellence in Education, 1983; National Science Board, 1983; Task Force on Education for Economic Growth, 1983; Twentieth Century Fund, 1983; however, for a dissenting view, see Shamos, 1988). To summarize the arguments made by advocates of science education: Not only will the economy require an increasing number of scientifically and technically trained professionals and support personnel, but most production and service jobs will require some quantitative and technical skills (Botkin et al., 1984; Education Commission of the States, 1982; but see Levin and Rumberger, 1983, for counter arguments). Moreover, an increasingly complex interlinking of the man-made and natural worlds makes it important for people to understand the basic parameters of both these worlds and their functioning so that individuals can make effective decisions in their personal lives and as citizens. Whether

the long-term goal of science education is scientific literacy for all or the development of science professionals or both, the short-term goals generally encompass three major categories of outcomes: knowledge, skills, and dispositions. The assessment challenge is how to probe student competencies in all three of these areas adequately and how to avoid certain adverse effects of testing.

Knowledge. The knowledge category includes knowing facts about the natural world -- knowing that the moon passes through cycles, that the shape of the leaves on trees in one's environment varies, that water droplets form on the underside of leaves on a humid summer morning. Also included in the knowledge category are the constructs (concepts), principles, laws, and theories that scientists use to explain why the moon appears to change shape, how leaf shape relates to a species' survival in a certain environment, and why the liquid droplets form on the underside of leaves. Gravity, heat, the Hardy-Weinberg principle, Newton's laws, and kinetic-molecular theory are examples of the theoretical knowledge scientists use to explain the natural world. (The set of organizing concepts identified by the Center's curriculum panel as integral to the elementary science curriculum, together with some teaching examples, is given in the Appendix.) Beyond facts about the natural world and the theoretical knowledge used to compose explanations for these facts, this category includes knowledge about the scientific enterprise -- its history, methods, philosophy, and values and its influence on human existence.

Skills. In addition to factual and conceptual knowledge, the goals of science education generally include three interrelated types of skills. Science laboratory skills are one type. The ability to read a thermometer, connect a wire to a terminal, stake out a quadrant, or focus a telescope are the skills that involve manipulation of equipment and observations of the kind required for doing natural science investigations. Another type is the set of intellectual skills called on in applying the methods of science. Among these are the ability to generate a hypothesis; to design an experiment that is a valid test of a hypothesis; and to collect, reduce, present, and analyze data (Frederiksen and Ward, 1978). The third skill type consists of generic thinking skills, including problem

solving and quantitative, logical, and analogical reasoning. These are component skillsof science intellectual skills as well as intellectual skills associated with other disciplines (Nickerson, 1988).

Dispositions. Acquiring a scientific knowledge base and developing the skills to apply the relevant knowledge to academic problems in the school setting are necessary but not sufficient. Unless science education also leads to the ability and inclination to apply science knowledge and science skills to new situations in one's work, in daily life, and in making personal and social decisions, neither the goal of developing productive science professionals nor the goal of scientific literacy for all citizens will achieved.

The Assessment Challenge

The valued outcomes of science education are varied, and each presents its unique assessment challenges. In general, knowledge is easier to assess in terms of time, effort, and resources than are skills or dispositions.

Assessing Knowledge. The first task in assessing the science knowledge acquired by students is deciding which categories of that knowledge are to be probed and what knowledge within each should be represented on a test. Once these decisions have been made, testing of factual and theoretical knowledge and knowledge about the scientific enterprise can be carried out with relative ease, using paper and pencil. This type of assessment format allows administration by a single person in group settings; hence, the exercises making up the assessment can be given to a large number of individuals. Because of the relative ease and efficiency of paper-and-pencil tests, particularly those -- like multiple choice -- that are machine scorable, most tests intended to provide national, state, or districtwide information on student achievement take this format (e.g., state-mandated tests, standardized tests available commercially, NAEP, IEA).

There is a second important characteristic of tests intended to assess science knowledge.

If the exercises are well constructed, the responses can be interpreted with reasonable certainty. A correct response indicates that the individual either knew the information required for the answer or was able to figure it out using information provided as part of the question. Determining the correctness of the response need not take into account the thinking skills the individual might have applied in comprehending the written item, in retrieving the fact from memory, in reasoning from the information in the item to the correct answer, or in eliminating incorrect responses. That is, the concern is neither with the means individuals may have used to access the information nor with the reasons for their conclusions; it is only with whether or not they have presented the correct information. In Table 1, we present some hypothetical illustrations of items testing factual science knowledge. The illustrations in this table and the ones that follow are in no way intended as exemplary test items; rather, they are meant to demonstrate that responses to factual items are relatively straightforward to interpret, whereas interpretation becomes increasingly more difficult for items intended to test skills and dispositions.

Table 1

Knowledge Assessment Exercises

Elementary

Which of the following best describes the path of the sun across the sky as it is observed in the United States?

a. east to west
b. west to east
c. north to south
d. south to north

Secondary

Which of the following is a statement of Newton's second law?

a. A particle not subjected to external forces remains at rest or moves with constant velocity.

b. If two particles interact, the force exerted by the first on the second is equal in magnitude and opposite in direction to the force exerted by the second particle on the first.

c. The acceleration of a particle is directly proportional to the external force acting on the particle and is inversely proportional to the mass of the particle.

d. Every two particles of matter in the universe attract each other with a force that acts along the line joining them, and has a magnitude proportional to the product of their masses and inversely proportional to the square of the distance between them.

Secondary

Which of these scientists lived at the same time?

a. Lavoisier and Lagrange
b. Franklin and Maxwell
c. Dalton and Bohr
d. Lyell and Wagonner

Assessing Practical Laboratory Skills, Science Intellectual Skills, and Generic Thinking Skills. The problems of skill assessment are much more complex than knowledge assessment. Assessing laboratory skills requires the use of laboratory equipment and introduces the distinction between knowing how to do something and having the competence to do it. To assess the latter rather than the former, assessment techniques need to be used that closely match the desired outcome, that is, the ability actually to carry out a given scientific procedure. This requires that, for assessment just as much as for instruction, students be provided with the necessary equipment. Assessment further requires that students demonstrate their capabilities as experienced observers evaluate and record their proficiency. Since the preferred method is labor-intensive and requires the use of materials, paper-and-pencil assessment is often substituted. Table 2 gives two examples but without the protocol needed to ensure appropriate observation and scoring.

Table 2

Laboratory Skills Assessment Exercises

Elementary

Measure the temperature of a liquid using a mercury thermometer marked in degrees Celsius.

Elementary

Find the mass of a metal block using an equal-arm balance.

Assessing the intellectual skills of science -- hypothesis generation, experimental design, data collection, and data interpretation -- introduces more confounding factors. Science-related intellectual skills are complex integrations of a variety of generic thinking skills with the ability to select and perform appropriate practical science laboratory skills. The first example in Table 3 illustrates an intellectual skill assessment exercise appropriate

for testing an elementary student's ability to design an experiment. The successful performer must know how to use a balance as well as have the logical skills to design an appropriate strategy to find the mass of the liquid apart from its container. In the design of most assessment instruments, science-related intellectual skills are assumed to be generic, skills that the student is expected to exhibit in any science context. However, not all testing experts agree with this assumption, arguing that familiarity with the context of the assessment exercise and the science knowledge available to the student are more important factors in the ability to perform an exercise than the science-related intellectual skills. It is certainly conceivable that a student could be successful possessing either science and context knowledge or science-related intellectual skills.

Table 3

Science-Related Intellectual Skills Assessment Exercises

Elementary

Find the mass of a sample of liquid contained in a beaker.

Secondary

Predict the relative quantities of heat required to raise the temperature of 100g of ice from -10° C to -5° C; from -4° C to 1° C; and 100g of water from 5° C to 10° C.

Design an experiment to test your prediction.

Perform the experiment.

Compare your results with your prediction.

Develop an explanation for any differences.

Design an experiment to test your hypothesis.

Interpretation of performance is extremely difficult for two reasons. First, not all observers will agree on what constitutes acceptable performance. And even if the observers do agree on acceptable performance, they may interpret the performance in different ways because the reasons for success or failure often are not evident from the responses. When a student performs well on the water/ice exercise in Table 3, it can be attributed either to familiarity with concepts related to heat or to the ability to apply generic science-related intellectual skills.

These same problems of exercise design and performance interpretation are presented in the assessment of the third group of skills -- generic thinking skills -- in even more severe degree. Table 4 gives a hypothetical example.

Table 4

Thinking Skills Assessment Exercise

Elementary

As a student who lives in North America, you observe the sun move across the sky from east to west. How would a student who lives in Australia, a country in the southern hemisphere, describe the motion of the sun across the sky?

a. west to east
b. east to west
c. north to south
d. south to north

The difficulty lies in interpreting the behavior that is elicited in an individual by an assessment exercise. When an individual performs an exercise and gives an answer, there is no way of knowing the mental processes and knowledge the person used in arriving at the answer. For example, if a person is given a description of a physical event and asked to explain it, a correct explanation may be the result of simply being

familiar with the situation and knowing the explanation for it. Alternatively, the person may be unfamiliar with the situation but recognize that a scientific principle he or she knows is applicable to the situation and apply the principle with the appropriate reasoning skills to come to a correct answer. A third possibility is that a person uses incorrect information in developing an explanation but uses a correct scientific principle and rules of logic to come to an incorrect answer. The exercise in Table 4 can be used to illustrate each of these possibilities. An observant student who has been to Australia may know the answer from direct observation or remember having read about sundials in the southern hemisphere. A student who knows the reasons for the sun's apparent motion across the sky and has the mental abilities to imagine how the sun's apparent motion across the sky would appear to a person in the southern hemisphere would come to the same answer but be using more sophisticated mental processes than the person who simply remembers. On the basis of the answer alone, the examiner cannot possibly know whether the performance represents recall; logical application of correct factual information, a scientific principle, and rules of logic; or right thinking with wrong information.

Obtaining information that sheds light on the methods and knowledge a student has brought to bear on the performance of an exercise requires individual administration of the exercise and collection of verbal protocols (Ericsson and Simon, 1984; Frederiksen et al., 1985; Nuthall and Lee, 1982). This method is highly labor-intensive. Moreover, test techniques that rely on verbal skills discriminate against children whose native language is not English. Even for native English speakers, there may be a confounding of verbal skills with science knowledge skills. In addition, one is never sure if the verbal protocol is a true reflection of how the answer was arrived at or a post-hoc explanation for how it might reasonably have been arrived at. Another difficulty with the use of verbal protocols is that different people interpret the same protocol or behavior in different ways. Another approach, provided the exercise involves several steps, is to track the development of the response step by step either by computer (Anderson et al., 1985; Brown and Burton, 1978) or by direct observation and subsequent interview. These procedures are costly and, like veral protocols, may suffer from difficulties of

interpretation and bias.

Not only are results of this kind of assessment difficult to evaluate, involving as they do interpretation of hands-on performance and of mental processes, they also are more difficult to report than those from a multiple-choice test. The data of real interest are qualitative rather than quantitative and not amenable to statistical tests or simple reporting.

Assessing Dispositions. Making judgments about a student's disposition to apply scientific knowledge and skills outside the formal classroom setting adds yet another level of complexity to assessment. One might attempt to assess disposition by the use of self-report -- that is, describing situations and asking individuals to indicate whether or not they would take a "scientific" approach to analyze them. This method has not yielded particularly trustworthy information (Gardner, 1975; Munby, 1983; Murnane and Raizen, 1988). A more appropriate method is to observe individuals to determine if they are scientific in their approaches to personal and civic problems. This method is resource-intensive, and even when attempted, the direct observations that result are difficult to interpret. Does failing to take a scientific approach indicate that the person has the inclination but not the requisite skills, the requisite skills but not the inclination, or neither? In addition, context has a profound influence on behavior. For example, not being scientific in approach in one situation might indicate that either the skills or the inclination is not in place. An alternative interpretation is that the person did not deem the scientific approach appropriate for that particular situation but would demonstrate inclination in other situations. Researchers have attempted to assess such proxy variables as impulsivity, attitude toward one's own competence, and fair-mindedness (Nickerson, 1988; Rowe, 1979), but much further work will have to be done before they can be linked with any confidence to the disposition to apply science knowledge and skills.

The Effects of Age and Experience. Age and its correlate, level of cognitive development, is another confounding factor in all science assessment. Performance on an exercise that indicates problemsolving for an 8-year old may well be recall of information for a 12-year-old. Moreover, the 12-year-old will be able to bring a greater wealth of experience to the exercise. Of course, the types of relevant experiences available to one youngster may be very different from those available to another who grows up in a different environment. For example, there is evidence that girls, even at an early age, have exposure different from boys to certain experiences relevant to the solving of some science problems -- fixing simple electrical or mechanical things, playing with motor-driven toys, building tree houses, using scientific equipment (Mullis and Jenkins, 1988). Since age is easily established, it can be factored into interpretations of assessment of performance, but the role of experience is difficult to take into account unless an assessment specifically collects relevant background information, as does NAEP (Hueftle, et al., 1983; Mullis and Jenkins, 1988).

Learning Over Time. The problems inherent in assessing complex learning outcomes can be analyzed in a more general fashion. In an article in the New Directions in Measurement series, Snow (1980) described a "continuum of referent generality" in both aptitude and achievement measurement. Referent generality refers roughly to the range of situations to which a given aptitude or achievement pertains. At the highest level, there might be aptitudes like general mental ability (IQ) or the kind of broad, complex, developed achievement measured by the SAT. At the lowest level, there might be aptitudes like speed of response time or achievements like two-column subtraction with borrowing. Important science learning outcomes are likely to be higher in referent generality than in narrower learning outcomes. Examples are students' understandings of scientific method or of such higher-level knowledge as the relationships between structure and function, the meaning of scale, or the concept of systems (see the Appendix and the Center's companion report on Curriculum and Instruction, Bybee et al., 1989).

Outcomes higher in referent generality are harder to teach directly because they must be visited time and time again, in a range of contexts, using different materials and different illustrations. They are harder to assess because they are less closely tied to any particular learning activity. The problem is how to assess understanding of the broad organizing principles, the inquiry approaches, and the ways of knowing that characterize science in the context of a particular learning unit given that these understandings may take years to develop. The problem is not unique to science, nor is it well solved in other content areas.

Erosion of Validity. Valid interpretation of test results may become more difficult as mandated assessments grow, particularly when they involve "high-stakes testing." The term refers to tests used to reach decisions that matter, where the stakes are high -- decisions about grades or placements of individual students, about teacher licensure and certification, or about rewards or sanctions (including public citation) for schools depending on their students' test scores.

Validity inheres not in a test itself but in an intended test interpretation, a score-based inference. There may be different logical bases for such inferences, calling for different strategies of test design and validation. Consider three examples: A college admissions test, a typing test for applicants for a secretarial position, and an achievement test administered by a state or district. The warrants for using the SAT or similar tests to help reach college admissions decisions include both logical arguments from the tests' content and design and empirical arguments from their observed correlations with college grades and other indicators of success. In contrast, the typing test directly samples a domain of performances that are a part of the work the person hired will be expected to do. The achievement test probably would be intermediate between these two examples: So far as it directly sampled some domain of proficiencies the children were expected to acquire, as use of a thermometer or an equal-arm balance, it would be like the typing test. So far as it was intended to show what children were likely to do or be capable of doing in nontest situations, its validity would have to rest on logical or empirical grounds -- areas that need much further exploration and work in the case of science tests (Frederiksen, 1986).

Erosion of validity may be said to occur when, as an indirect result of using the test, the warrant for the intended score-based inferences is weakened. In the case of college admissions tests, coaching that concentrates on test-taking skills or practice with feedback in answering multiple-choice items may improve test scores without bringing any concomitant improvement in the complex, developed aptitudes the test is intended to reflect. If such coaching improves the scores of some examinees, the correlation between test performance and subsequent college success is likely to be reduced, thereby eroding the test's validity as a predictor. (Of course, a longer-term program of coaching that focused on the underlying skills the test was intended to assess might improve both test performance and criterion performance. That would not affect the test's validity.) In the case of the typing test or reading a thermometer, it is more difficult to imagine any kind of training that would substantially improve test performance without also improving criterion performance. A work-sample test is highly resistant to erosion of validity.

Once more, an externally mandated achievement test would fall somewhere in between. Suppose that the mean scores for different schools were reported in the newspapers or used in other ways that created incentives for improving test performance. As with the SAT, scores would be likely to improve if children were given practice answering items similar to those on the test. (Teaching the particular items on the test itself would raise even more obvious questions of test score interpretation.) Probably few teachers would spend much time having children answer multiple-choice questions just to improve test-taking skills, but the more subtle influence of multiple-choice testing in many classrooms could well be increased use of worksheets, fill-in exercises, and question-and-answer recitations and diminished attention to activities bearing less resemblance to the tests such as extended writing, classroom discussions, or hands-on activities. These changes would focus instruction more narrowly on tested outcomes and thereby erode the validity of inferences from test performance to the full range of intended learning outcomes.

By the same token, the more closely test items resemble desirable instructional activities,

the less risk there is that even high-stakes testing will result in validity erosion. It might be argued that, if test exercises represented a proper balance of the full range of instructional activities, such erosion of validity could not occur -- teaching to the test would then be entirely appropriate. Unfortunately, no time-limited test could achieve such a mix. As noted earlier, some instructional activities are simply not amenable to that type of testing. However, assessment (although expensive, time-consuming assessment) could come much closer if it involved portfolios of students' work, systematic teacher observation, and other innovative strategies.

Of course, the idea of validity eroding implies that it is present in the first place. As discussed above, many achievement tests, including both teacher-controlled tests and externally mandated tests, support only very limited inferences to important learning outcomes because they test trivial facts or call for no more than low-level recall and rote problem solving. Whether these tests come to be influential because of the rewards or sanctions attached to them (high-stakes testing) or whether their influence arises through teachers' well-intentioned but misguided reliance on end-of-chapter tests in textbooks to define the goals of instruction, such "measurement-driven instruction" falls far short of the panel's and the Center's vision of effective and appropriate elementary science instruction.

Effects on Curriculum and Instruction. We have voiced the concern that the increased demands for testing, because of the characteristics of the tests generally used, will aggravate the tendency to reduce instructional activities to a set of measurable behaviors that pupils should demonstrate. The main curriculum effect will be to trivialize instruction by stressing, through lecture and reading assignments, bits of factual knowledge easy to test and likely to appear on most tests. Suppose, however, that a science assessment is created that is appropriately aligned with the sort of instruction that characterizes good science teaching, combines formal and informal assessment approaches (including performance on hands-on activities), and successfully addresses higher-order knowledge and skills as well as accurate and significant science information. Using such an assessment as a guide could dramatically improve the present state of most elementary science teaching. But if teachers themselves do not possess a firm

understanding of both science content and science curriculum goals, even the best of assessments will not be sufficient to guide their classroom instruction. Teaching must aim at the inculcation of knowledge, skills, and dispositions, not the replication of inventories of specific behaviors (Strike, 1982), and for that reason excellent teaching will always be more than mere imitation of excellent instructional activities. Teaching that aims only to reproduce correct responses may succeed in teaching manipulations of materials or rules for generating formulaic answers to formulaic questions without imparting any knowledge or understanding of value beyond the testing situation. To give an example, if teachers are told: "Activities involving wires, batteries, and bulbs will be used to assess exploration," there is a risk that all the children will soon manifest the particular behaviors to be elicited, but the understanding of what exploration in science means will remain as elusive as ever. Thus, even hands-on activities and laboratory experiments, considered the hall mark of good science teaching (Penick, 1983), run the danger of being reduced to a set of prescribed behaviors, leading to unreflective cookbook activities.

The right pedagogical move, the right question to ask or answer to give, depends on many particulars of the context and the learners. Sound assessment can provide signposts for instructional goals and desired student attainment to teachers, children, parents, and policymakers, but it cannot ensure their realization. We argue below that the reform of assessment has an important part to play in the improvement of science education, but it is not the entire solution.

Correspondence Between Curriculum, Assessment for Instruction, and Assessment for Monitoring

Should an assessment match the curriculum? For what type of assessment is this critically important? For what type of assessment should there be concern with more general goals of science education not necessarily tied to a specific curriculum? And if there is a close match between curriculum and assessment, will assessment results

provide a good indicator of the quality of the science education program and of the adequacy of student learning? (See Rudman et al., 1980.)

Matching Assessment to Curriculum

In considering the need for matching curriculum to assessment, the distinction between assessment for instruction and assessment for policy purposes becomes important. Clearly, if a teacher is interested in finding out how well students have learned a particular topic and set of concepts or have acquired competencies needed to perform certain science operations -- whether hands- on use of science tools or requisite reasoning skills -- the assessment should match as closely as possible the material that was to be learned. A related type of assessment with a somewhat different purpose concerns evaluation of curriculum quality. A curriculum developer or teacher trying out a new unit or laboratory exercise may be interested in how well it works by investigating whether students learn the intended material. In this case, also, the subject matter knowledge and competencies being probed by the assessment need to match closely those embedded in the curriculum material that was taught.

The case is somewhat different for assessments having a broader policy purpose. Administrators and policymakers may be more interested in the general level of knowledge and competencies that students have gained from their science instruction than in how well they learned from a specific curriculum. This is particularly true if the accomplishments of students in different countries or states or of students from different demographic groups (i.e., varying in ethnicity, socioeconomic status, or size of community) are to be compared to one another. Policymakers may also wish to compare achievement levels of students in the same location (say, the U.S.) over time, whether or not the curriculum might have changed in the meantime.

Broad-scale assessments that need to take into consideration the common core of curricula taught to all the students being tested (for example, NAEP as originally conceived, IEA, and assessments that use standardized, norm-referenced tests) will, by design, avoid special topics or concentration on subject matter taught to only a small fraction of the students being tested. This sets up a tension between the knowledge and competency students are able to demonstrate on a particular assessment and those they may have that the test does not probe.

It is at least conceivable that the inherent lack of correspondence between externally mandated large-scale assessments and specific school curricula and teacher-controlled tests will drive many such curricula (and concomitant teacher-controlled testing) toward the lowest common denominator, particularly if the large-scale tests are tied to policy consequences that affect individual schools, teachers, and students. The example of minimum competency testing stands as a warning of the potential for watering down the curriculum when attempts are made to set standards to be achieved by all students through the administration of a test that all are supposed to pass. A potential way of avoiding this danger is to institute multi stage testing (Bock and Mislevy, 1987), in which the level of each student's knowledge and competencies is established through a brief pretest, the results of which then dictate the difficulty of the rest of the items administered to the student.

A different approach, one that the proposed state-by-state NAEP mathematics assessment may be taking, is to assess the extent to which students have achieved prespecified, valued goals (say, in mathematics), irrespective of the extent to which the current curriculum reflects these goals. In the long run, this may have salutary effects on the curriculum, but in the short run, it is likely to yield dismayingly low test scores.

The Present Correspondence

Suppose the correspondence between curriculum and assessment is high. Does this

indicate a good state of affairs for science education? Teaching adequately to the three major goals of science education -- acquiring substantive factual and theoretical science knowledge, acquiring laboratory and thinking skills used in science, and developing the disposition to use the acquired knowledge and skills -- requires teaching for depth of understanding rather than for breadth of factual information. To achieve the desired depth, the teacher likely will want to introduce a variety of inquiry-based experiences for the children. In addition, the teacher will probably use some class time for group work and discussion among the children. If the environment is structured so that the children feel safe asking questions and clarifying their ideas, they will be able to use these activities and genuine discussions to build stronger and deeper understanding of science knowledge and stronger competencies needed to carry out inquiries. Obviously, these teaching strategies take time.

The goals for science education are not new. They have been expressed by many scientists and educators over the last fifty years. Yet the needed teaching strategies happen all too seldom in elementary school classrooms today. School science curricula, textbook publishers, and test makers have elected instead to promote breadth of coverage -- the learning of information consisting of a lot of small bits of knowledge and their rote applications to simple problems. Indeed, there is a kind of correspondence in place right now. Tests -- local, national, and international -- and textbooks from most publishers seem to be sending compatible messages (American Association for the Advancement of Science, 1985; 1986; 1989). It is important to know a little about a lot of things. Real understanding is not so highly valued. And science, though considered important in states where there is high-stakes science testing, is not as important as the basic skills.

Tests. Emphasis on recall of facts and formulaic problem solving unfortunately is characteristic of tests intended for teacher use, especially the end-of-the-unit quizzes and problem sections in textbooks, which match the breadth-of-coverage approach of the textbooks.

Tests designed for broad policy purposes generally show the following characteristics: A large number of mostly unrelated items are constructed for each of the major areas of science: life sciences, earth and space sciences, physical sciences, and scientific inquiry. For the reasons already discussed, the overwhelming preference is for the multiple-choice item. Occasionally, some tests will use a few written items with open-ended, short-response formats, and less frequently, some have attempted to assess students' competence in carrying out scientific tasks through using actual performance exercises. Unfortunately, the emphasis on factual recall matches the curriculum pretty well.

Textbooks. Not only do tests focus on breadth of coverage, but so do textbooks and teachers. It is hard to blame textbook publishers for wanting to sell their books in as many states as possible. They do this by scrutinizing curriculum guidelines from the majority of states (with particular concern for the states with the largest markets) and proceed to incorporate as many of the state curriculum objectives into their books as possible. Individual science curricula tend to be quite inclusive in their coverage of topics, and when curricula are taken together as a group, there is little in science that is not mentioned. The result is that books have gotten larger, and the number of topics included in a text has increased. By necessity, then, the number of pages devoted to any one topic has decreased, and the treatment of each topic has become more superficial.

Pressures on Teachers to Cover the Curriculum. Generally, if there is a state test used for monitoring and if there are high stakes associated with such a test (e.g., individual test scores or school performance are reported out), local school boards send clear messages that they want the students in their schools to do well. This means that teachers are expected to cover the topics in the district's curriculum and the state's test. Unfortunately, most elementary teachers are not experts in science. Therefore, whether or not there are pressures to teach toward a high-stakes test, the teachers' insecurity about their own knowledge of science causes them to rely on the textbook as "expert." The combination of the teachers' lack of understanding and the superficiality of the textbooks requires students to memorize words, facts, and "concept" statements that they (and often their teacher) do not really understand. "Learning science" in this

manner mimics rote memorization of vocabulary words and grammar rules from a foreign language one does not comprehend.

Science Has a Low Priority. Science is taught in a larger school context -- a context that clearly signals that what really matters are the basic skills. Teachers understandably spend more time on reading, mathematics, and writing than they do on science. Consequently, those science activities like hands-on experiments that require time for ordering materials, setting up, and cleaning up afterward tend to disappear from the curriculum. After all, time is short; besides, the conclusions of the experiment are generally presented in the textbook for students to read about.

Results of the Present Correspondence. The present state of affairs in many elementary schools is that science gets short shrift by teachers who are not terribly knowledgeable about science and do not have the confidence to engage in authentic inquiry together with their students. The textbook has become the science curriculum, and much of science learning is passive and superficial. This is reinforced by tests that largely assess factual recall and rote problem solving and rarely require a deep understanding of science. Is it surprising, then, that elementary students do not know much science?

Changing the Present Correspondence

Over the last five years, much interest has been expressed in changing the status quo (National Commission on Excellence in Education, 1983; National Science Board, 1983; Task Force on Education for Economic Growth, 1983; Twentieth Century Fund, 1983). Policymakers are not pleased that the nation's students do so poorly on national and international tests, particularly in view of the tremendous investment made in the schools. (See, for example, the statements made at the September 23, 1988, news conference on the results of the NAEP 1986 science assessments by members of Congress, the Assistant Secretary of Education, the presidents of the AFT and NEA, and representatives of the science and education communities. The statements are available from Educational Testing Service, Princeton, N.J.). Furthermore, one cannot

take refuge in the fact that U.S. students have skills different from students abroad for example, that they know more than students elsewhere but can't apply their knowledge, or conversely, that they have less science knowledge but are better problem solvers. It appears that U.S. students do not have a lot of science knowledge compared to students in other countries, nor are they better at solving the types of problems that appear on tests (International Association for the Evaluation of Educational Achievement, 1988).

Assessment as an Entry Point. We have pointed out that the use of poor tests is not the only factor exercising negative influence on science instruction. In most schools, teachers are under pressure to maintain an orderly and quiet classroom and to move through much material quickly so as to cover the textbook; therefore, they often feel the need to tidy up the messy business of science and get students to "get to the point," forgetting that false starts and off-beat ideas along the way are part of the point. Indeed, several fronts need to be addressed simultaneously if student achievement is to improve significantly. Certainly, better trained teachers who are given the opportunity to spend more time on teaching science would be one good place to start. But there is another place to start: with the premises that underlie the state, national, and international tests as well as the curriculum-embedded tests matching the current textbooks that so largely control today's elementary science curriculum.

The basic argument is that a critical entry point into breaking the present correspondence of mediocrity is to develop a different kind of assessment of science learning. In testing as well as in teaching, less may turn out to be more. Students should be able to demonstrate a deep understanding of science knowledge and the skills needed to do science. What these tests might look like is discussed in the next chapter.

The Uses of Assessment

There are many appropriate uses for valid assessment results. In the classroom context, teachers can use assessment to document growth across time and to determine the needs of individual students, based on their initial skills and gaps. These assessments can be

either diagnostic or evaluative. Policymakers may want to know the current health of the system and have answers to broad comparative questions in terms of improvements from the status quo and differential performance for various populations of interest (i.e., various demographically distinct populations and such geographic subunits as schools, districts, states, or even countries). The various participants in the educational enterprise need information to guide intelligent decision making about the next steps, whether these steps are at the microlevel of tutoring an individual student or at the macrolevel of recommending a change in high school graduation requirements. Assessment results can provide some of the input to such decisions (McLean, 1985).

The main distinction made here and elsewhere in this report is between assessment for instructional purposes and assessment for policy purposes. There is, however, another important distinction, roughly parallel, that is relevant to a discussion of the uses and misuses of assessment, namely, whether the testing is controlled by the teacher or externally mandated.

Teacher-Controlled Testing

Most problems in the use of assessment results arise in high-stakes testing situations where tests are externally mandated and test scores (or other outcome measures such as dropout rates) have direct policy consequences -- rewards or sanctions for schools, changes in curriculum or graduation requirements, placement and other career consequences for teachers. Classroom testing controlled by teachers and used for their own instructional purposes generally does not have any adverse policy consequences since the results are not generally shared with policymakers. There are, however, examples of testing used for instructional purposes that may indeed be ill-conceived. One such problem derives from instruction that is narrowly measurement driven, as exemplified by the Chicago Mastery Learning -- Reading system, which divided the K-8 reading curriculum into a sequence of 271 separate objectives to be mastered, and is reputed to have led to students spending so much time filling out worksheets that they never had time to read actual books or other meaningful materials. However, though

the testing was instituted for instructional rather than monitoring or accountability purposes, it was externally mandated by a large urban school system, as opposed to teacher-controlled testing.

It seems unlikely that this kind of measurement-driven instructional system will be applied to science in elementary school. It is conceivable, however, that teacher-controlled instructional testing might turn into a watered-down version of this sort of measurement-driven instruction. Teachers whose own comfort with science teaching and whose own level of scientific knowledge are low may look to tests for guidance about what and how to teach science. They may draw the unwarranted (invalid) inference from children's ability to answer low-level test questions that their students are achieving adequately in science. When confronted by complex, open-ended assessment questions, they may respond by direct teaching of possible responses, thereby subverting the validity of the assessment. Unfortunately, teachers with limited confidence in their ability to teach science are not uncommon at the elementary level: Weiss (1987) reports that only 27 percent and 15 percent, respectively, believe themselves to be well qualified to teach the life sciences and the physical or earth sciences, contrasted to 82 percent who consider themselves well qualified to teach reading. The perceptions of these teachers may be quite accurate since half of all elementary school teachers report never having had any inservice education in science, and most have had few if any science courses in college.

What might children learn from poorly constructed tests given by their teachers and invalid interpretations of their test performance? One unforeseen consequence or possible negative is that children may take the content of the test to represent science and therefore be turned off from any further study of the subject, concluding, for example, that science consists of nothing but a lot of facts and vocabulary words to learn. Children (and their parents) might also conclude from their poor performance on invalid tests that they lack the aptitude for science or scientific careers, that they "just can't get it."

The opposite problem could occur as well. For example, children might gain the impression from their elementary school experiences that science is a series of fun, informal activities where little is expected and all can succeed. Then, in middle/junior high school, students are suddenly confronted with an explicit science curriculum, difficult tests, a lot of mathematics, and perhaps a shared stereotype that science courses are too rigorous for girls or those not mathematically inclined. An important value of assessment in elementary science might be to let children know from the outset that there is subject matter to be learned in science, skills to be acquired, better and worse answers, and good and poor ideas -- that science is a curriculum area of the same kind as reading or arithmetic. Students also may come to understand through science assessments that science learning matters to their teachers and parents; indeed, perhaps serious assessment and reporting of science learning could make it important to parents and teachers.

Externally Mandated Assessments

The low incidence of science instruction in the nation's elementary schools complicates the uses of assessment mandated for policy and accountability purposes. Although well-conducted assessments can offer policymakers the information they need to make sensible decisions about programs and resource allocations, this is possible only if the broader context is understood as given assessment results are interpreted.

Time for Science in Elementary School. One of the most important contextual factors is the dearth of science instruction in elementary schools, as documented by recent studies. For example, in the 1986 science assessment conducted by the National Assessment of Educational Progress (NAEP), teachers of third-grade students were asked how much time they spent teaching science compared to carrying out other classroom activities. Approximately half the teachers at grade 3 reported spending one to two hours each week providing science instruction, and another 21 percent reported spending even less time than that (Mullis and Jenkins, 1988). Although questions can be raised about the validity of responses provided by third-graders, their reports agreed with those of their

teachers on the small amount of class time devoted to science instruction. Eleven percent of the third-graders reported never having a science lesson in school, and another 13 percent stated that they had science classes less than once each week. These NAEP data generally agree with the findings of the Report of the 1985-86 National Survey of Science and Mathematics Education (Weiss, 1987), in which elementary school teachers, K-3, reported spending an average of only 18 minutes per day teaching science -- less than half the time spent on mathematics instruction and one-quarter the time given to reading instruction in these early grades. The average amount of time spent on science instruction in grades 4-6 was 29 minutes, again less than the time spent on reading and mathematics. Further, these estimates had not changed from those provided by teachers in 1977 (Weiss, 1978).

Misuse of Test Results by Policymakers. The danger is that assessment will be incorporated into the nation's elementary schools either without the prerequisite attention required to increase and improve instruction or in ways so divorced from instruction as to render meaningless results. More and more, there is the temptation to test now and ask questions later. For example, national and international science assessments (International Association for the Evaluation of Educational Achievement, 1988; Mullis and Jenkins, 1988) have shown poor results for students in the United States. Given the current lack of instruction, these findings should not be surprising (Horn and Wallberg, 1984). However, without careful consideration of the appropriate use of assessment results, such negative findings may initiate a chain reaction that will foster unintended consequences rather than improved science instruction or achievement.

Although broad-based assessment for monitoring purposes is entirely legitimate, invalid inferences based on results from such assessments can lead policymakers to respond inappropriately. For instance, if the reasons for low achievement results are not well understood, the temptation may be to treat the symptoms rather than the underlying causes of the disease. Thus, it is considerably easier to focus legislation on increased numbers of courses to be taken or more student time to be spent on studying (e.g.,

recent reforms in the areas of reducing absenteeism, strengthening high school graduation requirements, and increasing homework) than to ensure that such legislation results in greater rigor in instruction or needed changes in assessment (Clune, 1989). In fact, it could be argued that, however well intended, legislative action that requires additional instruction without providing adequate resources for teaching subject matter content effectively may cause more harm than good. Requiring students to sit through extra hours of misguided instruction may lead to student indifference or exacerbate student dropout rates.

The Burden of Testing. Another consideration for policymakers is the amount of testing going on in any given classroom. Consider the following developments: NAEP is moving to state-representative assessments; other national or multinational studies are now focusing on or including science tests in their surveys (e.g., the NAEP six-country science study, NELS:88); states are increasing mandates for science assessments (from 13 in 1984 to 29 in 1987; see Blank and Espenshade, 1988) added to existing state reading, writing, and mathematics assessments, and local districts are following suit. Add to this that teachers, finding these large-scale tests unsuitable for their purposes, carry out their own testing programs to track student progress and assign grades. At least three problems are likely to arise as a consequence of the mounting number of tests given. Two of these have to do with resources consumed by testing that may be diverted from other purposes: first, a decrease in instructional time because of time given over to testing and, second, the increasing costs of test development, administration, and analysis (testing is now a billion dollar industry). The third problem has to do with potentially depressed test scores resulting from no real motivation or inclination to perform well on tests that have no personal consequences.

1. A decrease in instructional time because of increased testing can take two forms: the time required for the actual taking of the test(s) and the time teachers use to prepare students for the test(s) if this is different from the instruction they would provide otherwise. For tests intended to monitor achievement levels of a large number of students where individual scores are

not required, judicious sampling procedures and matrix administration of test questions (Messick et al., 1983) can, taken together, considerably reduce total student time consumed by testing. A potential drawback is that these testing methods work best when tests are centrally administered, requiring individual students to leave their classrooms and further disrupt their usual schedule. As to teachers taking time from their own instructional programs to prepare students for the test(s), this may become a growing problem as test scores are used to reward or censure individuals or schools. However, if tests could be developed and used -- despite the abundant difficulties discussed earlier -- that come close to representing all the important goals of science education, teaching to the test(s) would become good instructional policy.

2. Developing good tests is a labor-intensive activity likely to keep teachers, administrators, and science and testing experts preoccupied while possibly displacing effort and money that might better have gone into curriculum planning, staff development, and other needed improvements. An alternative though not necessarily a money-saving one, is to use commercially available standardized tests for monitoring and accountability purposes, but these tests are unlikely to mirror the goals and curriculum of the state or the district and therefore will provide only a general indicator of low-level skills. This is why states that are giving strong curriculum guidance, like California, are also investing considerable effort and resources in constructing assessments that match their curricular goals. Whatever the investment at the test development or test purchase end, there have to be resources invested at the other end in analyzing the results, interpreting them, and reporting them. When state or district testing is mandated without sufficient funds for analysis or reporting, as is sometimes the case, the testing itself may be a waste of money. In any case, the cost of all these functions associated with testing adds to the administrative expenditures of schooling and may pull resources away from direct instruction.

3. A different sort of problem is the concern that students may not take seriously

tests that have no consequences for them personally -- for example, neither grades nor college admissions depend on the test results, individual test scores are not revealed. Anecdotes by teachers on student attitudes and behaviors while taking districtwide or statewide tests give some substance to this concern. We know of no systematic study that has analyzed patterns of test item responses from this perspective, but the effect is well known in survey research. When questionnaires become too burdensome for respondents, they may not complete them or give only perfunctory responses (Bradburn, 1979; Sharp and Frankel, 1983; Sinaiko and Broedling, 1976). If this phenomenon is equally real for test responses, it is likely to grow as students are required to take an increasing number of tests for monitoring and accountability purposes.

Validity of Assessments for Policy Use. Beyond broad legislative reforms that may or may not change for the better what happens in an individual school, there are the effects of high-stakes assessments discussed earlier. When policy actions that affect individual schools, administrators, teachers, or students are taken on the basis of assessment results, assessments become very important. Teachers are more likely to teach to the test, and it is naive to ask them to avoid doing so. Thus, the issue from an assessment perspective is to improve the quality of such tests so as to make instruction based on their content worthwhile. To achieve this goal, assessment developers and administrators need to evaluate the validity of tests in light of the following criteria, as should interpreters and users of assessment results. (More specific questions that speak to these criteria and that should be asked of science tests are given in Chapter III.)

1. **Ecological validity.** Does the test measure what educators care about? An earlier section of this chapter points out the discrepancy between such curricular goals of elementary science education as increased proficiency in science tool use and in thinking skills that are difficult and costly to measure and the overreliance on multiple-choice tests that are easy to administer and score. Before embarking on assessment for monitoring or instructional purposes, the goals of science education should be articulated and the

assessment instrument(s) examined to determine if, indeed, the questions asked in the assessment reflect these goals.

2. **Correct science content.** Do the test items or assessment exercises represent good science? A recent study by the National Academy of Sciences (Murnane and Raizen, 1988) found that 5 to 10 percent of items on each of nine commonly used science tests included inaccurate or misleading science statements that decreased the usefulness of the test results. If an item is poorly written, students may give the wrong answers for the right reasons (or for reasons unrelated to science learning) or because an item's content is erroneous or misleading. For example, Hein (1987) points out that students who understood the scientific principles of ice melting missed this item on a standardized test because the graphs were plotted inaccurately.

3. **Reflecting science accurately.** The content and format of a test sends a message to students about how educators view the subject being tested. Most science tests could readily be construed to mirror science as a wealth of dry, elaborate, and unconnected details to be memorized by rote. Further, if students are made to practice material resembling such test items either in preparation for districtwide tests or in form of the quizzes appearing at the end of sections in their textbooks, the message that science is boring, obscure, and irrelevant is reinforced.

4. **Cognitive style.** In addition to being dull and uninspiring, tests may require thinking or reasoning that is antithetical to scientific habits of mind. For example, in the absence of knowing the right answer, random guessing may be a good test-taking strategy for some kinds of tests; conversely, thinking hard about a problem may lead a student to question the "right" answer but give it anyway because that will increase his or her test score. Moreover, tests may inadvertently reflect the cognitive style stereotypically associated with the sciences. There is some evidence that males and females may frame questions

somewhat differently if no less rigorously (Cohen, 1987), and that one's cultural background can influence understanding about the relationship between humankind and nature.

In short, if the assessment does not measure valuable content or if it contains errors, the results should not be used for policy or instructional decision-making. Such results are invalid from the perspective of measuring what elementary students have learned about science, and any interpretations based on them are likely to be faulty. Interpretations will also be flawed if the referents used to judge student learning are based on national norms created to rank students rather than on the development of competencies important in understanding and doing science. Actions based on these misinterpretations risk being inappropriate if not harmful, squandering resources or encouraging bad teaching.

Assessing Science Programs

Educators and people concerned with educational policy have multiple goals for elementary science education, yet they rely on a narrow and limited set of outcome measures to assess its quality. If more comprehensive and alternative measures of students' science learning were to be developed and used, as urged above, considerable gains for students, teachers, and policymakers would be achieved. But good science education is not limited strictly to outcomes. Educators, parents, and policymakers are also concerned about the quality of children's day-to-day science experiences. This is one reason assessments of science education should include measures of what schools are able to provide (e.g., whether they have time allocated and materials designed for hands-on, inquiry-oriented science; whether school norms press children toward high achievement and aspirations in science; and whether science teaching is provided by individuals who are qualified, committed, enthusiastic, and energetic). Only with the inclusion of such measures can science assessments be used to understand the quality of children's science experiences in school.

These features of school science programs also are important because program characteristics shape students' learning outcomes in complex ways. For example, as Barr and Dreeben (1985) have made elegantly clear, classroom experiences and interactions are at the very heart of the educational enterprise. These can be linked to student outcomes with some confidence. And, since classroom experiences take place within a particular school, their quality is affected by the characteristics of the school. More precisely, school characteristics create conditions that enable or constrain science teaching and learning.

Assessments of school science programs, then, can provide information about central features of science education -- features important to observe in order to learn more about the circumstances in which particular outcomes are produced. If one neglects to consider school program characteristics as important mediators of inputs from outside the school (e.g., resources, state policies, and local district policies) and as influences on classroom experience (and, through these, children's science learning), an over simplistic portrayal of science education will result.

Obviously, assessments of school science programs cannot possibly provide the complex data researchers need to understand fully the relationships among program characteristics and science outcomes. However, they can provide useful clues to policymakers about problem areas and strengths. The challenge is to design assessments that provide the most central information with the least number of indicators.

III. ASSESSMENT OF STUDENT LEARNING

What to Assess

We have stated that valid assessment requires a clear definition of the goals and contents of the curriculum: subject matter to be taught and the skills and competencies students are expected to acquire. It should also consider what is known about how children learn science.

Curriculum Content

For science in elementary school, consensus has been building on the goals, content, and nature of the curriculum. In the preceding chapter, we noted that science education should concern itself with three aspects of learning and understanding science: knowing important facts and constructs of science; gaining skills that characterize the doing of science, including laboratory skills, skills needed in applying science methods, and generic thinking skills; and acquiring the dispositions that incline individuals to apply the knowledge and skills they have acquired to new situations. But defining these critical aspects of science learning is not enough. The knowledge and skills and dispositions must be embedded in subject matter through which they are to be taught, and the choice of subject matter depends on its centrality to each field of science.

In parallel with our report on assessment, the Center has developed a Report on Curriculum and Instruction. This report suggests that the elementary school science curriculum be organized around nine major concepts -- powerful explanatory constructs that are applicable to science and technology and beyond and that accommodate different developmental levels. The report defines these concepts and provides several examples of appropriate teaching topics for each. To provide some substantive reference for the succeeding discussion of assessment of student learning, we list the concepts here: organization (or orderliness), cause and effect, systems, scale, models, change, structure and function, variations (discontinuous and continuous properties), and

diversity. Short discussions and teaching examples that illustrate appropriate topics for lower and upper elementary school for each of the organizing concepts are given in the Appendix.

Learning Science

In addition to addressing the goals and objectives of science education, the curriculum and associated instructional strategies must consider what is known about how children learn science. Cognitive scientists working in collaboration with scientists from the fields of science and mathematics generally taught in school have shed light on several areas relevant to teaching science and assessing science learning (Resnick, 1987).

The Child as a Maker of Theories. Studies in the areas of mathematics and science have demonstrated that students construct their own views about how numbers behave and how the natural world works, views that they bring to the classroom (Anderson and Smith, 1983; Gentner and Gentner, 1983; McDermott, 1984; Stevens et al., 1979). These views, however, do not necessarily correspond to the laws of mathematics and science being taught in the classroom. Erlwanger's (1975) case study of Benny, a very bright, motivated, and successful elementary school student, demonstrates how even the best students carry around misunderstandings about numbers. Similar findings have been made for several physical phenomena, as summarized in Murnane and Raizen (1988:59).

Students don't just "outgrow" the views they have formed. College students graduating from well-rated institutions and many highly educated adults hold conceptions of natural laws that are at odds with scientific constructs. It takes patient elucidation over time and opportunities for students to surface their self-constructed theories so that they can test them against evidence (Driver and Oldham, 1986). If understanding is the goal, then students need opportunities to develop that understanding through an accumulation of knowledge gained from a combination of direct experience and knowledge from experts (including the textbook and the teacher) and to consider whether their own beliefs gained from prior experience are consistent with their new experiences afforded

through classroom activities, with their own lines of inquiry, and with the canonical explanations of scientists (Champagne et al., 1982).

If the world is dissonant with their beliefs, students begin the process of constructing new understandings and beliefs. This is not something that happens in ten or fifteen minutes. It takes time. It takes several trials. It takes talking with one's peers and the teacher and consulting evidence provided in formal compilations. It requires multiple observations, carefully done, and checking one's results against those of others. It takes developing a disposition of open-mindedness and being willing to change one's mind in the light of new evidence. This kind of learning takes place in a risk-free environment where not knowing is considered the first step to acquiring knowledge.

Elementary school teachers who want to encourage an environment of inquiry have a considerable advantage over teachers of older students. For one thing, young children are innately curious; they enter school with hundreds of questions about how things work. Second, they don't mind getting their hands dirty; they will happily "mess around" with water and sand and animals and chemicals. The teacher does not have to worry about developing intellectual curiosity, rather, how to protect and nurture it; how to prevent it from drying up and disappearing in an environment in which all too often science education doesn't begin with the child's beliefs and questions about the world but with a list of technical words to learn and, later, formulaic applications of "scientific laws." The implications for assessment are clear enough: Current tests must be changed if they are not to reinforce the most sterile of science teaching.

Solving Problems and Higher-Order Thinking. Another line of research that has implications for science learning and assessment has contrasted problem solving by experts to problem solving by novices. Researchers have inferred that, given a problem situation, "experts" in the area bring to bear a highly organized knowledge base that allows them to see patterns and relationships not obvious to the novice and thus to solve the problem efficiently (Larkin et al., 1980). In fact, what may be a difficult and novel problem for the novice may be a routine one for the expert. Of importance in

improving science instruction and assessment is understanding the structure of the knowledge base that the expert brings to bear and how individuals come to acquire and build such a knowledge base in a specific subject area. Does the close observation of a given phenomenon over many days or even years, under different conditions, enable them to understand the universals inherent in the phenomenon as contrasted to the surface features? Does this then allow experts to categorize a new problem and relate it to the phenomena they know in a way that generates efficient solution approaches? Does deep understanding of one area allow one to think metaphorically in other areas that do not seem related on inspection of surface characteristics only?

Resnick (1987:3) has described some features of higher-order thinking that characterize problem solving in science (as well as in other fields):

- Higher-order thinking is nonalgorithmic. That is, the path of action is not fully specified in advance.
- Higher-order thinking tends to be complex. The total path is not "visible" (mentally speaking) from any single vantage point.
- Higher-order thinking often yields multiple solutions, each with costs and benefits, rather than unique solutions.
- Higher-order thinking involves nuanced judgment and interpretation.
- Higher-order thinking involves the application of multiple criteria, which sometimes conflict with one another.
- Higher-order thinking often involves uncertainty. Not everything that bears on the task at hand is known.
- Higher-order thinking involves self-regulation of the thinking process. We do not recognize higher order thinking in an individual when someone else "calls the plays" at every step.
- Higher-order thinking involves imposing meaning, finding structure in apparent disorder.
- Higher-order thinking is effortful. There is considerable mental work involved in the kinds of elaborations and judgments required.

Research studies on problem-solving and higher-order thinking, as well as research on the knowledge children bring to the science classroom, point to the need to pursue a given topic or phenomenon in depth. Murnane and Raizen (1988:125) state:

> At present, there is an emerging literature that relates the depth of coverage of subject matter to student understanding of the content (Glaser, 1984; Sizer, 1984). Deeper, more complex coverage of a concept or set of concepts increases the opportunity for students to be engaged in effective complex problem solving (Chi et. al., 1981; Resnick, 1987). Not surprisingly, these researchers have also found that people's capacity to understand and remember new information in an area is related to their prior level of understanding of the area, and that experts in a field approach the solution of problems differently and more efficiently than do novices. This discussion suggests that the depth of coverage of material in a curriculum is an important aspect of its quality

Problem Solving and Collaboration. Scientific inquiry in the real world is seldom done in isolation. Students working as problem solvers should have the experience of working in small teams as well as individually. They should be able to collaborate on developing approaches and question one and other's interpretations, testing individual ideas against those of others in the group. In this way, students are able to sharpen their communication skills in the context of working with a real problem related to a scientific phenomenon. Cooperative/collaborative learning could be used at least part of the time for observing scientific phenomena, solving multistep problems, and designing and conducting experiments. This implies that testing, as well, should at times probe the work of student groups.

An apt illustration is the "paper-towel test" carried out by students at the Shady Hill School in Cambridge, Massachusetts (personal communication, Sally Crissman, August 15, 1988). A class of fifth-graders was asked by a fictitious restaurant owner to recommend the best brand of paper towel to use in his restaurant. The children, working in groups, designed and carried out various tests: price, taking into account the cost of a single roll, the number of sheets per roll, sheet length, ply, and area per roll; absorbency (mls of water absorbed per sheet); dry strength (number of rubs per towel); and wet strength (grams supported by a wet towel). They recorded and graphed their

findings and wrote their recommendations based on interpretations of their data, to the restaurant owner. The group reports exhibited a considerable range of quality and depth. The teacher was able to glean much additional information by watching how the groups of students worked, made decisions, resolved problems of methodology, and so on. She made these observations part of the assessment record by noting them in her journal.

Implications. Several clear messages emerge from this deepening understanding of how effective science learning takes place. The first message is that:

Less is more.

Effective science teaching takes time. It is important for students to be able to ask genuine questions, conduct genuine inquiry, and be guided to find answers and not let the teacher be the only question-asker in the classroom. To make this kind of science teaching possible, the curriculum needs to concentrate on a few areas deeply rather than on a lot of areas superficially.

Does a commitment to depth define the "ideal curriculum"? Unfortunately, it does not. Developing a framework with cogent examples -- the Center's or an alternative -- is still necessary. But even as substantive choices are made, a commitment to depth does send a message about the characteristics of a good curriculum. It changes the emphasis from one concerned with spanning a large domain of knowledge to one concerned with deep understanding. It argues for thorough treatment of whatever is studied and for providing every opportunity for students to deepen their understanding of scientific constructs appropriate to their level, and to hone the laboratory and thinking skills that will allow them to pursue science questions with increasing rigor.

A second message concerns the role of the learner in the educational process:

Responsibility for learning is shared between learner and teacher.

As Murnane and Raizen (1988:74) summarize:

> Recent research in cognitive science (Resnick, 1983) and the growing acceptance of generative or constructionist psychology (Osborne and Wittrock, 1983; Watts and Gilbert, 1983) further highlight the importance of the student in the learning process. The current view of the student learner is one who actively constructs his or her own meaning, rather than serving as a passive receptacle of the teacher's transmitted information. The constructionist's view of the learner places great importance on the prior knowledge of the student and the nature of the learning activities in which the student engages. Because learners have some control over the nature and quality of their efforts, some of the responsibility for learning outcomes shifts from the teacher to the student.

A third message deals with the relationship between learning factual knowledge and developing higher-order thinking skills:

Different types of learning are not hierarchical; the acquisition of facts and structuring of a knowledge base goes hand in hand with learning how to apply knowledge, how to reason and solve problems.

There are important implications for the assessment of science learning in these messages. Assessment, if it is to reflect what students are expected to have learned, needs to be grounded in the intended curriculum and in the instruction that precedes the assessment. Further, if the development of higher-order thinking skills is an important goal of the science curriculum, the requisite effort must be invested to create assessment exercises and strategies that truly probe for these skills, and the time must be taken for adequate administration of the new forms of tests and for the analysis and

reporting of results.

How to Assess

What kinds of assessments would foster curricula and instruction that focus on science understanding and development of the tool use and thinking skills characteristic of science? Before addressing this question, we summarize various points made earlier on the current state of testing. One thing is certain: Fundamental changes are needed in both classroom testing and broad-scale assessments.

Testing Today

The present state of testing, in the classroom and out, is discouraging. Resnick (1987:34) holds that "most current tests favor students who have acquired lots of factual knowledge and do little to assess either the coherence and utility of that knowledge or the students' ability to use it to reason, solve problems, and the like." She points out that, if high test scores are the objective, such tests will decrease emphasis on the teaching of higher-order skills and, instead of continued use of such tests, she calls for assessments that rather than fixed answers will require techniques that themselves depend on judgment and that are open to alternative interpretations." More broadly, she concludes that assessment alternatives must be developed that are more suited to the goal of teaching higher-order thinking (Resnick, 1987:47).

Broad-Scale Testing. Educators in school districts that have been recognized for the excellence of their K-6 science programs (Penick, 1983) also have expressed concern about current tests and the extent to which they assess what students are learning in inquiry-based, hands-on science classrooms. A summary of a recent conference on elementary science education that discussed innovative programs agreed that standardized achievement tests are not adequate for assessing what elementary students learn in good science programs and urged development of improved tests and alternative evaluation techniques (National Science Resources Center, 1986).

As noted, the great attraction of multiple-choice testing is that this format represents an economical means for assessing extent of factual knowledge since test responses can be scored rapidly, reliably, and relatively cheaply (Murnane and Raizen, 1988). The efficiency and reliability of all multiple-choice tests becomes particularly attractive for large-scale assessments. Whether all students are to be tested (as in some statewide or districtwide assessments) or results from a representative sample of students are to be generalized over a large population (as in NAEP or IEA), time constraints, response burden, and costs of test administration and analysis drive assessments toward traditional formats.

Teacher-Made Tests. Unfortunately, the use of multiple-choice or short-answer test formats that can be scored objectively is not limited to the sort of large-scale assessments where alternative forms of testing are most difficult to carry out. Bloom, (1984:13) writes that "teacher-made tests (and standardized tests) are largely tests of ....remembered information it is estimated that over ninety percent of test questions the U.S. public school students are now expected to answer deal with little more than information. Our instructional material, our classroom teaching methods, and our testing methods rarely rise above the lowest category of the [Bloom] taxonomy -- knowledge."

A recent study (Dorr-Bremme and Herman, 1986) found that teacher-made tests, together with teacher observations and judgments, play a large role in influencing what happens to a student. Externally mandated tests are not unimportant since they are often used for initial student placement. But more critical are the techniques used by the classroom teacher to assess student achievement and performance because they govern a variety of decisions that impinge directly on students -- what curriculum sequences individuals will be exposed to; the educational experiences they will have; assignment to classrooms; chances at further education; and grades and related information reported to parents, prospective employers, and colleges and universities. The authors conclude that "the various teacher-designed strategies of achievement

assessment cumulatively shape students' learning environment, academic self-concept, educational status, and (ultimately) their socio-economic opportunities (p. 104)."

Yet, despite the importance of the assessments carried out by them, teachers are hardly prepared for this critical function. Studies going back 20 years (Ebel, 1967) and carried out more recently (Fleming and Chambers, 1983) have documented the inadequacies of teacher-made tests: a variety of commonly occurring errors, a preponderance of short-answer questions that emphasize memorization of facts, and a lack of questions that require knowledge application or other higher-order thinking skills. These findings are to be expected considering the poor preparation that teachers receive in this area. For the most part, neither their undergraduate education nor their practice teaching/internships, nor even subsequent in-service programs treat testing and assessment skills as an important competency that teachers need to acquire. (Coffman, 1983; Rudman et al., 1980; Woellner, 1979; Yeh et al., 1981). For example, a recent survey (Dorr-Bremme and Herman, 1986:105) found that only about one fifth of the teachers responding "received staff development related to selection and construction of good tests or in use of test results to improve instruction."

This sort of information on the quality of tests and other assessment strategies constructed and used by teachers for their own purposes is disquieting. It appears that these tests and assessments are no better at probing highly valued but hard-to-assess outcomes of science education than the short-answer tests constructed for efficient use with large numbers of students. Of course, teachers who are insecure in their knowledge of science and lack experience teaching it are likely to feel the need to stick closely to textbook facts. These teachers can hardly be expected to develop imaginative science tests, no matter what their training in general testing skills. Nevertheless, we agree with Dorr-Bremme and Herman (1986:105-106) that "it seems worth considering just how qualified today's teachers are to be developers of the tests that most affect students' lives. How effective are teacher generated tests in revealing insufficiencies in individual students' learning? How valid are they as measures of student achievement? How do teachers decide how often to test? How skilled are elementary school teachers at

analyzing the commercial curriculum embedded tests that they frequently use? Similar questions can also be raised about teacher skills in making observation- and interaction-based judgments of children's learning."

Characteristics of Assessments of the Future

In our view, an assessment of science learning, whether for use by the classroom teacher, by a district or state, or at the national policy level, is authentic only if it matches the curricular and instructional goals of science education as they have been briefly outlined above. What would such assessments look like?

- Assessments would match exemplary instruction. Assessment exercises would be indistinguishable from good instructional tasks.
- Exercises would include hands-on performance tasks to allow students to demonstrate their proficiencies in laboratory and science thinking skills.
- Assessments would strive to probe the child's depth of understanding as well as mastery of a body of knowledge.
- The emphasis would be on both the approach and the product, on how an answer was obtained or a hands-on activity carried out, and on the "correctness" of that answer or performance.

In Great Britain, current reforms in assessment are designed to address both the need for monitoring and the improvement of instruction in an assessment approach that should be considered in this country. The various strategies to be employed include both formal and informal means of assessment, for example, notes kept by teachers on their observations of student discussions, profiles of student performance over time, structured exercises and examinations, and standardized tasks administered to students one on one or in small groups (Department of Education and Science and the Welsh Office, 1987). In addition, teachers are encouraged to arrange for students to demonstrate to outside audiences -- the PTA, other teachers and student groups, school authorities -- what they have learned and can do in science. The use of mixed assessment strategies accompanied by quite specific teacher training represents a serious

effort to incorporate the features we suggest for assessments of the future.

Assessments That Match Instruction. Recently, several examples of the British approach to assessment have been published. Lock and Davies (1987:277-279) describe the Oxford Certificate of Educational Achievement (OCEA) philosophy used to assess students from ages 11 to 16. The teaching, learning, and assessment of science are closely interwoven activities, and the teacher may switch from one to the other as the need arises. For example, when a student demonstrates inability to carry out a specific laboratory task during an assessment, the teacher may want to turn to instruction and defer completion of the assessment until the student has remedied this deficiency. To illustrate in the context of the science class example that introduces this report:

> *After several weeks of working with seeds, Ms. Lopez wants to assess whether the children have developed good ideas about testing factors that are important in plant germination and growth. As the children set up various conditions, she notes that some are having difficulty weighing soil and solid fertilizer and measuring water and liquid fertilizer. She takes time out to work with the children on these skills until they have mastered them.*

Ms. Lopez uses assessment in a formative manner to shape instruction, as do teachers in the OCEA scheme. Evaluations of student performance are based on evidence drawn from several sources -- direct observation, discussions with students, and written work. Obviously, all these are also important aspects of teaching and learning. Another dimension of this approach is that, like instruction, assessment is not limited to one point in time; students have several opportunities to demonstrate their knowledge and proficiencies and can do so in different contexts, analogous to the way good instruction proceeds.

Hands-On Work. Assessment and testing procedures designed in the context of the British work on Assessment of Performance Units (1984-1985) have emphasized performance on "practical" (or hands-on) tasks. Incorporating such tasks in assessment procedures highlights the importance of the laboratory and application component of school science and, it is hoped, will influence the science curriculum to include more of

these activities. Woolnough and Allsop (1985) suggest that three roles for laboratory and applied work are valid:

(a) Developing practical skills and techniques
(b) Problem-solving in a scientific way
(c) Developing a feel for phenomena

The reason for the emphasis on hands-on activities and, later on, more formalized laboratory is the interplay between factual knowledge, understanding of scientific constructs, and practical work. Carrick (1987) notes that "observation is greatly influenced by the conceptual framework of the observer. At the same time, experience of practical work helps pupils to understand what they are learning. According to Woolnough and Allsop they build up 'tacit knowledge' as well as more formal or 'explicit understanding'." For example, the children in Ms. Lopez's class could study seed dispersion through illustrations in a book. But by collecting seeds from the environments where they occur naturally and by observing them blow and flutter in the wind, adhere to clothes, scatter on the ground, and be eaten by birds, children are able to develop their own experiential knowledge of seed distribution and matching dispersion structures.

The development of practical skills and techniques and of problem solving in a scientific way can be probed through appropriate assessment tasks. As in tests of factual knowledge, however, it will be important to guard against allowing assessment of hands-on activities to become trivial and purposeless. Rather, patterns of performance assessment of practical and science thinking skills need to correspond to the best practice in teaching these skills.

Probing the Student's Understanding. Prior knowledge that students bring to science instruction can facilitate or impede further learning. For example,

> *before beginning the seeds unit, Ms. Lopez had invited the class to talk about seeds by asking: "What are some examples of seeds?" The children were eager to contribute ideas and called out: beans, acorns, nuts, corn nuts, raisins, radishes, poppy seeds, potatoes, flower seeds, peanuts, and other answers. All the answers were listed on the*

board. The teacher mentally noted which of the children hung back -- some might be shy, some might be unsure of what a seed is. Later, she would find other, unobtrusive ways to probe the understanding of some of these children individually. For the class as a whole, Ms. Lopez was gratified by the predictable enthusiasm and interest the children showed. Though they didn't know just why she asked them such a question, they were happy to go along.

She noted something interesting about the set of answers they had given: the children seemed to associate seeds with things they ate. She thought she could build on that when she was ready to introduce the idea of seeds as having concentrated energy to help plants grow. The children also seemed to associate seeds with plants. But their conception of seeds had limitations. The teacher made a mental note to bring a coconut to school, a pine cone, and some peppercorns. She wisely refrained from the temptation to add any suggestions of her own to the list at this time, but the diversity of seeds was something the class needed to work on.

When the 15-minute activity was finished, Ms. Lopez had done several things. She knew what the children understood in a general way, and she had some hints about which ones might require some extra help later on. She also perceived which children seemed to have a lot of ideas, and later, when the children would work in small groups, she would try to have each group include one of these children. She had identified some "hooks" to the children's own understandings and experiences that she could capitalize on later. All this, and the children didn't even know it was an "assessment"! At the end of the day, Ms. Lopez jotted down some informal observations on index cards for many of the children. When the time came to assign grades, these informal notes would be important.

Students' beliefs about physical phenomena often can be discovered by asking students to draw or otherwise indicate what they thought was happening or would happen under certain conditions. Teachers sometimes mistakenly believe that, because their students provide correct answers to multiple-choice tests, they understand the scientific explanations underlying the phenomena in question. A well-designed test assessing depth of understanding would provide opportunities for students' beliefs that differ from canonical scientific knowledge to surface through the use of probes asking students to explain what they were thinking when they gave certain explanations (Almy and Genishi, 1979). The knowledge being gained about students' prior beliefs -- the areas in which students hold on to their experiential knowledge in the face of instruction that provides explicit canonical scientific explanations -- should provide a rich source of assessment questions. (See Helm and Novak, 1983, for several volumes of studies on scientific

"misconceptions" in a variety of areas.)

Two very different examples may prove illustrative. The first of these concerns a computer simulation unit called ThinkerTools developed by White and Horwitz (1987) to enable children to understand the laws of motion. The instructional techniques were designed to facilitate four key stages of knowledge acquisition: motivation, inductive learning, abstraction, and transfer. A computer-generated, simplified microworld was created as a way to teach sixth graders the basic constructs and laws of Newtonian mechanics, a major area of tenaciously held "misconceptions." The students had science class every school day for 45 minutes, and the ThinkerTools curriculum occupied the entire class period. The curriculum took two months to complete. The evaluation included a 13-item transfer test of the underlying principles to real-world contexts. Students in the ThinkerTools curriculum, who had been able to test their beliefs against the evidence provided by the microworld, averaged 11.2; students in the control group averaged 7.6. Short interviews with students demonstrated considerable differences in depth of understanding of Newtonian mechanics between the ThinkerTools students and the control group.

The work by Champagne et al. (1980:10-11) also holds promise for interesting assessment strategies. Two types of tasks were developed: the DOE (Demonstrate, Observe, and Explain) task and the Con SAT (Concept Structuring Analysis Technique) task. The work also included systematic observation of students as they planned, executed, and analyzed experiments. In each case, the task is not a stand alone assessment exercise; the subject matter for the task is chosen in conjunction with the science topic being taught at the time. The main purpose is to gain information on the knowledge base students bring to the science topic, how this knowledge base is structured, and how students apply it.

The DOE tasks were administered in a group setting. In the DOE task, students were asked to predict the result of a demonstration and the basis on which the prediction was made. Students recorded what they observed in the demonstration and noted any

inconsistency between the prediction and the observation, attempting to explain how they might resolve the inconsistency.

The ConSAT task as originally designed provided information about a student's understanding of the technical terms of mechanics, specifically their definitions and relations to each other. The ConSAT task was administered individually. The student was given a stack of cards, each with one term on it that a physicist might use to describe the motion of an object (kinematics) or to explain the cause of observed motion (dynamics). Sixteen terms were used: mass, weight, volume, density, object, time, distance, speed, position, velocity, acceleration, force, pressure, work, energy, power. The students read each term aloud and decided if they recognized it. If they did, they were asked to define the term. Unrecognized terms were set aside. When all terms were sorted and the recognized terms defined, students arranged the recognized terms on a large sheet of paper in a way that showed how they "think about them." When students completed the arrangement of the terms, they were asked to explain why the terms were arranged in that way and to specify the relations among individual terms or groups of terms. Finally, students reviewed the unrecognized terms. Any term that they now recognized was defined and placed in the structure. It seems quite conceivable to adapt both this task and the DOE tasks for use in assessment.

The following might represent a concept structuring task for Ms. Lopez's seed unit:

Materials: (The specimens for this task can be collected by students as they do activities on seeds.)

1. A set of cards or plastic envelopes each with an intact seed and a dissected seed with parts attached. Dissected parts should include the seed coat and cotyledons.

2. A set of cards or plastic envelopes consisting of subsets composed of groups of four. Each subset includes a seed, a plant, a blossom, and a fruiting body from a single species. The sets preferably would contain dried specimens, but pictures are satisfactory.

3. A set of cards consisting of subsets of cards in pairs. Each pair includes a card showing an animal and a seed that it eats or transports.

The assessment task is for the student to select cards or envelopes from the individual sets or combinations of the three sets and arrange them in a way that shows something the student has learned about seeds.

Possible arrangements include:

- Sorting seeds into groups according to
  - size
  - color
  - mode of dispersion
  - structural adaptation for dispersion
    - wings
    - hooks
    - fluff
    - weight
    - digestibility of the seed coat
  - where collected

- Sorting plant parts according to species and arranging plants and parts to show the reproductive cycle

- Matching structural adaptations of the seeds with the animals that disperse them (for example, a cherry matched with a blue jay and a photograph of bird spore with an intact cherry seed).

Two important dimensions of assessments designed to probe the depth of students' understandings in science are time and the type of answer that is acceptable. Appropriate assessment questions or exercises would largely be based on the kinds of understandings that students are expected to have developed after sustained exposure to a scientific domain, including opportunities to collect evidence and question their own beliefs. That clearly implies that some assessment activities would take place over a longer time than that of a typical class test. Students should have many opportunities to produce appropriate behaviors so as to enable them to use feedback (both self-feedback and that received from the assessor) to refine their performances.

Probing for depth of understanding also means asking essential questions -- ones that seek to get at the core of a discipline. In science, that implies asking some questions that may have multiple solution paths and more than one answer and posing disorderly

situations where "the problem to be solved" is not prespecified and students may have to conduct their own investigations.

Attending to the Process of Problem Solving. Carey and Shavelson (1988) point out the importance of tracking the process by which a student obtains an answer to a problem, because how the answer was derived may be more important than whether the "correct" answer was given. Several attempts at a solution may be necessary before a successful one is found, and often this trial process is more significant than the mechanical application of formulaic solutions that often characterize multiple-choice achievement tests. Real-world problems, argue the authors, are difficult because they require representation of the problem, goal setting, and planning the solutions, sometimes repeating these steps several times and comparing alternative approaches. Relatively routine substitutions of numbers into formulae represent the last and sometimes easiest step in arriving at an answer. Therefore, "problem solving steps and the conceptions underlying them should be assessed more fully and efficiently than at present because of their importance in mathematics and science activities (p. 213)."

Both in Britain and in this country, assessment strategies have been advocated that allow students (and teachers whose students take part in large-scale assessments) some choice in the problems to be addressed. The element of choice has important ramifications for the depth/breadth issue discussed above. Giving students (or teachers) a choice in the problems a student will be asked to solve acknowledges the fact that not everything has to be covered. A premium is placed on depth of coverage -- on problems with many parts (some well structured, others badly structured) that require depth of understanding.

Attention also should be paid to the process by which a child sets up and solves a problem. In traditional tests, students are given the problem and asked to solve it. Sometimes the most difficult part is "setting up the problem," weeding out the extraneous information and figuring out what problem needs to be solved and what information is needed to solve it. Recent work in dynamic assessment (Campione and

Brown, 1987; Feuerstein et al., 1987) suggests that, if a child is having difficulty setting up a problem, the assessor could have available a series of probes and questions that might help a student determine how to approach the problem. In this sort of procedure, the role of the teacher or assessor is to ascertain how much help the child needs before he or she can solve a problem. The instructor provides the minimal amount of scaffolding necessary for the student to be able to solve problems. This process is also described by Collins et al. (in press) in a recent paper on cognitive scaffolding.

Two less important but still significant aspects of problem solving should also be assessed from time to time: the degree of precision a child uses and whether a child checks his or her work. Again, multiple assessment strategies and several different assessments are appropriate as students build competence in these skills.

Criteria for Choosing Tests and Exercises

We have tried to establish the premises and portray the philosophy and spirit that should guide formal assessment, whether conducted for purposes of improving instruction or monitoring performance. What do these premises and philosophy imply for the selection of assessment exercises and instruments? In the next section, we elaborate on the four general criteria given in Chapter II for evaluating tests. We illustrate these criteria with specific questions distilled from the preceding discussion, through which teachers and principals may want to screen tests.

Questions to Ask about Tests. We concern ourselves here not with the psychometric properties of tests but with their substance. The first six of the following questions were originally suggested by Akers (1984:34-35 as quoted in Shavelson et al., 1988:149) for mathematics textbooks. They are equally appropriate for science curriculum materials and science tests; in fact, these sorts of questions might well be asked about the classroom instruction provided in science classes as well.

1. Are there problems that require students to think about and analyze situations? Akers suggests that as an alternative to word problems that require simple computations, textbooks (read "tests") should include some thought problems. For example, Ms. Lopez's students, who are also learning addition and subtraction, might be asked to evaluate statements according to whether they make sense, for example, "we have 25 seeds; 12 are different and 15 are duplicates" or "we have 8 different kinds of seeds; 4 kinds are eaten by people, and 6 kinds are eaten by birds."

2. Does the test feature sets of problems that call for more than one step in arriving at a solution?

3. Are problems with more than one correct solution included?

4. Are there opportunities for students to use their own data and create their own problems?

5. Are students encouraged to use a variety of approaches to solve a problem? For example, Ms. Lopez might ask at the initiation of the seed unit: "Where can we look for seeds to bring to class?"

6. Are there assessment exercises that encourage students to estimate their answers and to check their results?

For science specifically, we would add:

7. Is the science information given in the problem story and elicited in the answer accurate?

8. Is there opportunity for assessing skills (both in the use of science tools and in science thinking) through some exercises calling for hands-on activities?

9. Are there exercises included in the overall assessment strategy that need to be carried out over time?

10. Are there problems with purposely missing or mistaken information that ask students to find the errors or critique the way the problem is set up? (What is wrong? What is difficult?)

11. Are there opportunities for students to make up their own questions/problems or designs (for example, design a seed that has more than one dispersal feature)?

We invite readers to make up their own examples of seed unit problems and assessment exercises that illustrate some of these 11 characteristics.

If teachers have insufficient time to create or search for acceptable assessment problems and exercises, at the very least, they can look at the end-of-chapter quizzes or the tests at the back of the book and use them as the starting point for assessing in some different modes. Multiple-choice questions can be converted to open-ended questions. Moreover, the questions in the books could serve as the basis for essay questions, discussions/ conversations, drawings or other representations of ideas, and the development by children and teachers alike of more interesting problems following some of the suggestions made above.

Improving Informal Assessment. Much -- probably most -- of the information teachers use to guide their instructional decision making comes not from formal tests but from informal classroom observations. There are ways of doing such observations better and, at the same time, increasing the credibility of these observations as a source of information for marking and grading, communicating with parents, and so on. First, these observations should be somewhat systematic, that is done regularly. Teachers might carry around a packet of index cards to jot down observations on what particular students do from time to time. They might spend a few minutes at the end of each day (at the very least, every few days) to file those observations for future retrieval. Teachers should be alerted to the human tendency to note the atypical and neglect the commonplace (Almy and Genishi, 1979). Routine observations are of value. It is also important that informal observations systematically cover all the children in the classroom. In short, teachers should be scientific observers.

> *After the "seed walk," Ms. Lopez assessed the differences in the children's understandings and eagerness to talk about the seeds they had collected. She noted which children easily made observations and which ones had more difficulty. She kept track of which children made the more obvious statements that she had anticipated and which ones came up with unusual or unexpected responses. She noted which children seemed comfortable using the lens for examining their seeds and which ones seemed more awkward.*

As she planned the group activities for the next day, Ms. Lopez used her notes to place children in groups of twos and threes. Her goal was to group children so that they would prompt one another's inquiry. She put shy children with more talkative ones. She paired children who seemed more skillful at making observations with those who had more trouble. She put children who seemed certain of their statements with those who asked difficult questions. She organized a "seed journal" activity for those children who said they wanted to work with their seeds by themselves.

In short, Ms. Lopez used her assessment of the differences among children to form groups that would work together most productively on the explorations and observations that would come next.

Informal observations on the face of it seem more valid -- less artificial and contrived -- than more formal, written measures. Unfortunately, they may also, on the face of it, appear less reliable. Reliability comes through replication. Multiple-choice tests are reliable in part because they are standardized across learners, but also in part because they involve the summation of many independent pieces of information, the responses to the many items. The same principle can be used to enhance the reliability and the status of informal observations. By aggregating over multiple occasions, reliability can be increased. Validity will be highest when such multiple observations also involve a degree of "convergence" or "triangulation," a synthesis of evidence from different contexts, employing different modes of representation.

If a child makes drawings illustrating an idea, talks about it, sets up a relevant experiment, and the like, the teacher's confidence can be high that the scientific construct or principle has been assimilated. To the extent that science is integrated into the curriculum, with common themes carried across content areas, opportunities for such convergent validation increase. For example, a unit on astronomy can be tied to the early explorers' use of celestial navigation; a unit on weather or geography can be tied to social studies; graphs or word problems in mathematics can be tied to regularities in natural phenomena observed during science study. It is important that students, parents, other teachers, school administrators, and public officials understand criteria that govern assessments using informal means. OCEA (Lock and Davies, 1987), for example, "does not remove the teacher's personal viewpoint, [but] it does place the assessment criteria

in the public domain. Both teacher and student are employing the same set of rules and these may be displayed on the laboratory wall for all to see."

Both formal and informal assessment strategies need to be part of an ongoing process that, over time, provides a profile of a student's progress within a grade and throughout the years of school. Taken as a whole, therefore, the various forms of assessment should provide information on a student's science knowledge and on the competencies acquired -- both mental and in the use of science tools -- to design and carry out inquiries in science. Obviously, not any one test will be able to address all these knowledge and skills competencies adequately, which is why we urge multiple assessment approaches.

Using Assessments in Elementary Science Education

Using Information from Assessments

We concentrate here on the use of assessment that matters most -- how to make science instruction more effective in the classroom. We conceive assessment to be a continuum serving formative and summative purposes, using methods that range from the informal to the formal.

> *Ongoing monitoring to find out what students know and the ability to use this monitoring as a basis for shaping instruction is woven throughout Ms. Lopez's instructional activities about the seeds -- at the individual, group, and class levels. For example, individual students were asked to keep journals -- ongoing records not only of student ability to make observations and communicate information but of growth in concepts and understanding about seeds. In addition to each student's written record, the "What We Know About Seeds" chart was updated at regular intervals. Thus, after each activity, the students were encouraged to add to the chart -- not only a variety of facts about seeds but understandings related to the nine organizing concepts that structure the elementary science curriculum in Ms. Lopez's school. For example, even initially after bringing in the seeds and surveying the class collection, students might have noticed that there are many different kinds of seeds (diversity) or that sometimes it is hard to tell what "is" and "isn't" a seed (organization) or that seeds grow into plants (change, systems).*

After their walk, they may have enhanced their understanding of the complexity of how seeds might be organized by discussing any differences between the seeds they found on the walk and the ones they had found where they lived. Such discussions of location, size, transportability -- "inside fruit or shells" or "blown by the wind" as compared to "on bagels" or "those that stuck to the socks" -- could also lead to understandings about structure and function as well as sharpening awareness about observations, the range of "data" being collected about the seeds, and how the data might be organized.

Finally, Ms. Lopez had many additional opportunities for evaluating growth of understanding for individual students or groups of students. Each student gave an oral presentation about a seed he or she had found and examined, participated in a group that was asked to write, graph, or confer with her, and shared in whole-class activities where each student question or statement was an indication of progress (or lack of progress). Ms. Lopez was interested in several different types of understanding: Did the children develop an understanding of the role of seeds and their various properties in propagating plants and in providing food for animals and humans? More important, did they develop an understanding of some of the nine principles that had been illustrated through the study of seeds? Were they more adept at using a lens and at measuring length, weight, and volume? And did they develop some sense of systematic observation, recording, and analysis of data as they collected seeds, organized their collection, and germinated the seeds? As Ms. Lopez kept notes on the progress of individual children and the class as a whole, she developed the source material that would enable her to make more formal assessments to be reported in report cards, to parents, and -- for the class as a whole -- to Mr. Sandowski, the 3rd grade teacher.

Short-Term Assessment. In the context of elementary science, ongoing formative evaluation is integral to good instruction and may comprise the bulk of the assessment activity -- particularly in the primary grades. Such assessment is important before, during, and after instructional units.

Before Instruction. As illustrated earlier, assessment preceding a particular area of study can be useful to determine what students may already know about the content and skills involved. Understanding students' prior knowledge both establishes the range of current understanding among students and provides a context for ensuing instruction. For example, in beginning a unit of study on weather, teachers may first ask students what words they know to describe weather phenomena. Such a brainstorming session can inform the teacher and motivate the children, particularly if used in conjunction with

effective questioning strategies that require students to make predictions and explore relationships among important constructs (see the DOE task described above).

Discussions that focus on students' existing state of understanding can inform instruction in several ways. As noted, students prior beliefs may be revealed and subsequently used to determine where to begin instruction, what kinds of activities to select, and what facts and constructs to emphasize. Teachers can also learn more about what types of instructional strategies may be required (e.g., hands-on, discussion, or investigative projects). Such preassessment may also indicate how best to group students for collaborative science activities so that their points of view and skills can be productively shared. Assessment before instruction may also be useful in establishing how much students know about the tools and ways of thinking that characterize science. For example, if half the students have never used a microscope or a particular measuring device, some extra help may be required before the equipment is used in a classroom activity.

During Instruction. As pointed out, formative assessment is also vital during instruction to monitor the success of particular activities and diagnose the needs and progress of individual students. Through observation and questioning, teachers can learn about students' understanding of major constructs and principles. By systematically observing students using tools and carrying out investigations, teachers can determine both whether students are able to use equipment and whether they understand what scientific inquiry entails. By asking probing questions, listening to small-group discussions, having students write up observations or procedures, or listening to brief oral presentations by students, teachers can see if prior beliefs have been replaced by more complete understandings and if students can apply understandings to new situations. If it seems that students' understandings are inadequate, then teachers may need to devise alternative activities or strategies to complement or reinforce earlier instruction.

Other informal assessment techniques can involve the use of diaries and journals kept by students. These approaches not only are invaluable instructional tools that help students

learn how to communicate in the field of science but also are useful to teachers as a way of monitoring students' evolving understanding.

Continual monitoring of student progress is crucial to successful instruction. If students are left to do only independent work and are made solely responsible for evaluating their own learning -- as in reading the textbook on their own and answering the embedded questions -- the teacher has no way of gauging actual student understanding in time to make adjustments in instruction. (This is quite aside from the problem already discussed that, by and large, text-embedded questions generally do not reinforce major principles and relationships but focus on relatively low-level knowledge and skills.)

After Instruction. Both formative and summative assessment are important after the teacher completes a unit of study. Formative assessment through discussing ideas, sharing group projects, or listening to oral reports can help students summarize what they know as well as encourage them toward further avenues of study and investigation. The key is for teachers to emphasize additional applications of major constructs, principles, and relationships to determine whether students have gained understanding or are simply regurgitating content in a rote fashion. If the latter seems to be prevalent, additional instruction may be advisable before implementing more formal evaluation techniques.

The role of summative evaluation and feedback after a unit of study may expand in the upper elementary grades as students become more proficient in both written and oral communication. Teachers may use more formal assessment techniques to give individual students feedback about their particular strengths and weaknesses as well as to fulfill the obligations of giving grades and pointing out progress for parents. However, summative evaluation at the end of the units should not revert to asking students to work exercises requiring rote recall any more than should on-going assessment. For example, teachers can use a portfolio approach whereby students' efforts throughout the unit of study are examined to assess their growth. Projects and research reports are also useful strategies to assess student learning. Whatever the method of assessment, students should be

asked to communicate what they have learned, and not simply fill in blanks or circle choices. Allowing students to communicate what they know helps reinforce understandings and gives students practice in important literacy skills. If students cannot communicate what they have learned, it is doubtful they have gained understanding. In addition, communicating results and knowledge is integral to the nature of science.

End-of-unit assessments should focus not on disparate facts but on important constructs, principles, and relationships that are critical in determining how best to initiate future units of study. If students do not grasp the essence of the unit, they are missing the foundation necessary for subsequent study. It bears repeating here that assessment sends a message to students about the nature of science. If assessment does not encourage good scientific thinking and incorporate the approaches taken in science to collecting and interpreting evidence, students will come to understand that what "really counts" in science learning is memorizing trivia. Such a message can only serve to dampen interest and detract from future motivation, no matter how sound the instruction. The innovative, recursive nature of assessment and instruction must be sustained through the summative stages of evaluation by using end-of-unit assessment techniques that reflect the thinking skills and the applications of important constructs and relationships stressed in instruction.

Long-Term Assessment. Four important functions of long-term assessment of science learning are: to monitor cumulative learning, to provide teachers and students with opportunities to engage in self-assessment, to guide program development, and to provide evidence to school boards and parents that demonstrates program effectiveness.

Certain learning outcomes -- problem-solving skills and quality of laboratory reports, for example -- occur in such small increments that assessment in short periods does not produce any discernible change in performance. Assessment over longer periods is necessary to monitor the development of these skills in individual students and the effectiveness of programs both at the school and at the subject matter level. Developing

problem-solving and written-communication skills are schoolwide objectives, and monitoring their development provides important information about the overall success of the school's program. Information about the development of these skills collected in different content areas provides information about the relative contributions of different subject areas to achievement of the schoolwide objectives. Analyses of such data provide the opportunity for teachers across subject areas to coordinate their teaching of these vital skills. Data for this type of assessment can be amassed by keeping portfolios of student products, provided that explicit standards are made public and observed. In the case of written communication, each student might be required to add a piece of written work from each subject area to his or her writing portfolio each month.

Beyond its value in monitoring the quality of the various curriculum content areas and of the schoolwide program, such a collection of written products is valuable in helping students assess their own progress. A conference with a writing teacher, where the student's work at the beginning of the year is contrasted with that at the end, is a valuable opportunity for the student to learn and practice criteria used in assessing his or her performance (in this instance, criteria used in assessing writing quality). This sort of information is important in the development of self-assessment skills.

Data from end-of-year assessments also provide important information about year-to-year articulation in subject areas. Assessment of year-end achievement matched against prerequisites for the following year's program provides teachers the chance to coordinate the school's instructional program within subject area and to increase the probability that students will be successful learners.

In addition to usual end-of-year achievement tests and portfolios, public presentations of year-end accomplishments are useful mechanisms for long-term assessment. The assessment exercise in this case is for a class to develop an end-of-year report chronicling the year's activities in science class and a student eye-view of what was learned. Such an activity serves as a motivator for record keeping throughout the year and gives students an opportunity to try to fit the individual pieces of the year's learning

together. Teachers, observing what their students retain of the year's experiences and what they have made of it, have valuable information on which to base a self-evaluation of their teaching performance.

The public presentation of the year-end report to parents and the school board gives students an opportunity to practice presentation skills while serving the important added function of informing parents and school board members about the quality and accomplishments of the program. For this method to be effective, the end-of-year report must be truly a class effort, which means that, at the public presentation, any student is prepared to deliver any part of the report.

Assessing Attitudes and Dispositions

Continuing Engagement with Science

In Chapter II, we pointed out that the disposition to apply science knowledge and science skills to new situations is a valued outcome of science education. We also noted the near impossibility of assessing this outcome.

Two different types of approaches might be examined for possible development of proxies to assess the inclination to apply scientific habits of mind outside the formalities of the classroom. One comes from the recent literature on critical thinking, which is replete with discussions about the dispositions of critical thinkers. A review paper by Baron (1987) summarizes several clusters of dispositions that were noted by leading cognitive psychologists, philosophers, and educators as important for effective thinkers to display. They include:

- Intellectual curiosity and independence
- Open-mindedness and objectivity
- Sensitivity and empathy
- Deliberation and reflection
- Metacognition and self-criticism

- Thoroughness, persistence, and precision

The second approach emphasizes the methods of doing science and the beliefs that make up the ethics of science (Welch, 1984). Generally included are (Blosser, 1984; Murnane and Raizen, 1988; Rowe, 1979):

- Objectivity and skepticism
- Tentativeness and flexibility
- Curiosity
- Commitment and perseverance
- Self-confidence in one's ability to do science

Listing these sorts of attributes does not particularly ease the assessment problem. Recent recommendations on assessment of science education have either specifically cautioned against measuring attitudes (Department of Education and Science and the Welsh Office, 1987) or placed a lower priority on them than on measuring student competencies (Council of Chief State School Officers, 1984; Murnane and Raizen, 1988; Shavelson et al., 1987). There are several reasons for this: attitudinal outcomes are generally of less interest to policymakers; their direct measurement and subsequent interpretation of results are fraught with difficulty (Raizen and Murnane, 1985); and there is inherent danger in their use on all but a highly aggregated basis, yet such high levels of aggregation wash out the very classroom effects that are probably important in engendering the attitudes being assessed. The only countervailing argument is that the very attempt to measure the attributes that characterize critical thinking and scientific habits of mind emphasizes their importance for students and teachers and may lead to attention being given to them in the classroom.

A possible way around these dilemmas is to measure observable student behaviors for example, interest in voluntarily undertaking science activities beyond prescribed classroom work (and subsequent enrollment in science electives), students' self-monitoring of their work, and monitoring of peers. Ms. Lopez might add observations on these behaviors to the records she keeps on her students. Conceivably, some structured performance tasks might also provide opportunity for observing these

behaviors, particularly if the tasks call for sustained work. At this stage of understanding, however, much more research is needed to identify student behaviors that are reliable indicators of future willingness to continue an engagement with science and the disposition to apply one's science knowledge and skills.

Attitudes About Science

Many assessments have included measures of attitudes about science, for example, liking of science lessons or science teachers, valuing of science as a contributor to society, plans for future science careers (Hueftle et al., 1983; Mullis and Jenkins, 1988). These sorts of attitude measures have two kinds of problems: (1) results are often paradoxical (e.g. "I like my science teacher" but, from the same student, "Science class is boring") and difficult to make sense of (Munby, 1983) and (2) the linkages between attitudes about science -- even if they could be better assessed -- and student achievement, let alone later dispositions to engage with and use science knowledge and skills, are open to question (Willson, 1983).

Equity Issues

Despite the difficulties of assessing dispositions and attitudes, there may be merit in asking a very selective set of questions that can provide information on equity issues. For example, it would be important to ascertain whether there are any systematic differences between feelings of efficacy among subgroups -- males and females, Whites, Blacks, Hispanics, and so forth. Assessing belief in the ability to do science in elementary school would make it possible to determine at what age any differences between subgroups begin. Knowledge about such differences could help emphasize academic press for science achievement (see Chapter IV) for the very groups who currently achieve poorly.

Another relevant example concerns beliefs about the pertinence of science careers. Even though responses to queries about future career plans are notoriously unreliable,

and the more so at earlier ages, it is important for all potentially science-able students -- girls, Blacks, and Hispanics included -- to believe that science careers are appropriate for them. Present data indicate that this is not so (Fullilove, 1987; Harvard Education Letter, 1988; Mullis and Jenkins, 1988), a condition that needs to be remedied given the demographic changes in store for the future work force (Hodgkinson, 1985).

IV. ASSESSMENT OF PROGRAM FEATURES

Why Assess Elementary School Science Programs?

Assessing key features of the science program is, as of now, an essential part of monitoring elementary school science education. At least in the short run, policymakers and educators need information about school resources, organizational characteristics, and classroom processes so that changes will be made on the basis of an accurate understanding of school conditions and undesirable changes will be avoided.

Three reasons underlie this argument. First, many policymakers, educators, and parents place a high value on the quality of the resources, people, and activities that constitute children's day-to-day science experiences. Thus, assessment of these characteristics has an inherent value. Second, because current methods allow measurement of only a small range of the learning outcomes in science, excluding some that are most highly valued, assessing program features may prevent schools from placing undue emphasis on "looking good" on the limited outcome measures that are available and narrowing their educational programs to do so. Third, even though understanding is limited on how programs produce the desired learning outcomes in science, information about science programs may provide clues about the context in which these learning outcomes come about. Such information can contribute important information to the political discussion about how to improve science programs (Oakes, in press).

Balancing the Effects of Assessment on Science Programs

The U.S. Department of Education, the National Science Foundation, the Council of Chief State School Officers, and nearly all individual states have efforts underway to identify educational indicators. At the federal level, indicators are seen as essential for monitoring the status of the nation's educational system and tracking changes over time. At the state level, policymakers hope that indicators will provide information that can be

used to hold local districts and schools accountable for their performance and to suggest directions for improvement.

We have noted that the very existence of external assessment systems like those currently being developed will influence how schools operate, and that these effects are particularly strong when "high-stakes" decisions are linked to assessment results -- decisions about student promotion, teacher evaluation, resource allocation, or school certification, for example. The importance of this point for the improvement of science education is driven home by work recently done for the U.S. Department of Education's Office of Research and Improvement on the development of state accountability systems. As a part of that work, a survey of the states by the Council of Chief State School Officers found that most states are moving rapidly to implement educational accountability systems, and that the centerpiece indicators of most of these systems are scores on standardized tests of basic knowledge and skills (U.S. Department of Education, 1988). The press on schools is particularly great since most state accountability data will be made public in disaggregations at the school or district level, and in many states, rewards and sanctions will follow from scores that districts or schools obtain.

Not surprisingly, schools in these states are marshalling substantial efforts to look good on the indicators. In other words, standardized testing programs are shaping the nature of the school curriculum and the learning experiences that schools emphasize. For example, data from principals and teachers in six states that RAND studied through the National Center for Policy Research in Education (CPRE) make it clear that these indicators will be a powerful force (U.S. Department of Education, 1988). Whatever else schools and teachers want to accomplish instructionally, high-stakes test-score indicators substantially affect teaching and learning that takes place in the classroom, and teachers spend a good deal of their energy and time attempting to raise students' scores.

Building assessment systems that circumvent these unintended and undesirable consequences should be a primary objective. In the long term, the development of measures that assess the full range of what students know and can do should have the effect of driving the curriculum in a positive way, and do so reasonably unobtrusively. Positive since, if outcome assessments include such elements as students' problem-solving skills and performance in hands-on application of science, instruction will also focus on these things. Positive also because, unlike assessments that measure directly whether and how particular classroom resources and processes are being used, it leaves decisions about how to organize and conduct instruction to those at the school site. Consequently, it does not tamper with teacher professionalism; moreover, it ensures that educators be held responsible for the bottom line of students' learning.

In the short term, however, the proclivity to use narrow test-score indicators for making decisions about science programs needs to be counteracted with equally influential indicators of valued science program features. Since adequate measures of the science outcomes that are valued most are not widely available, they are rarely used to press schools to develop programs that emphasize all these outcomes. School programs can be assessed to establish whether they include the time, materials, teaching resources, and other attributes likely to enable unmeasured but desired learning to take place. In the best case, such assessment systems should support schools' and teachers' emphasis on those program characteristics that appear to support students' development of a sophisticated understanding of science constructs, performance, and critical thinking in science, and general problem-solving skills -- outcomes not currently measured well. In the short term, then, program assessments may provide the best hope for righting the understandable but unhealthy tilting of school programs toward low-level knowledge and skills. It may be the most reasonable way to use assessment to exercise leverage over the quality of science education -- at least until better outcome measures are developed.

Program assessments are needed for other reasons. First, program measures can enhance the usefulness of assessments by permitting analysts to desegregate outcome data by important subgroups. This disaggregation will permit a better understanding of outcome trends. Relevant subgroups include more than the conventional divisions of students by race, class, gender, school locale, even though these are extremely important for understanding the distribution of science outcomes. Disaggregations of data by subgroups of students who have experienced similar school programs are also of interest. Identification of these subgroups is not possible unless program characteristics are assessed as well as outcomes.

Recent data from the International Association for the Evaluation of Educational Achievement's (IEA) Second International Mathematics Study demonstrate this point nicely. First, the study collected data about the type of classroom 8th graders were enrolled in (e.g., remedial, typical, enriched, or algebra). It also collected information about whether tested 12th graders were enrolled in calculus courses or were at a "pre-calculus" level. Desegregating outcome data into these enrollment-related subgroups produced patterns suggesting that it was the students enrolled in the "lower" class levels who accounted for a substantial portion of the relatively low achievement levels of U.S. students in comparison with those in other nations (McKnight et al., 1987).

Program measures can also permit analysts to generate clues about why subgroup outcomes are what they are. Again, the Second International Mathematics Study analyses are illustrative. In addition to collecting data about the classes in which students were enrolled, the study collected "opportunity-to-learn" information; that is, the study queried teachers about whether their students were provided instruction in the topics represented in test items. Thus, when data about enrollment-based subgroups were analyzed, it was also possible to observe that U.S. students in the lower-level 8th grade classes and the pre-calculus 12th grade classes had significantly less exposure to the topics and skills that were tested than did students at the same grades in other

countries or their peers in higher-level classes in the U.S. The juxtaposition of these data suggested that one clue to the lower level of achievement of these students was that they had not been taught the material (McKnight et al., 1987).

Data such as these in science could enable policymakers and educators to pinpoint areas in science education that may be problematic (e.g., the lack of opportunity of many U.S. students to learn particular constructs) and to target their reform efforts more precisely.

What Program Characteristics Should Be Assessed?

Deciding which program characteristics to include in an assessment system poses problems given the limited understanding of which features are most central to the quality of students' science experience or which function as the most important mediators between school resources and outcomes. Additionally, many program characteristics that are highly valued and are believed to affect students' understanding and interest in science lie beyond current measurement technology.

Nonetheless, the literature on science education and schooling generally provides several clues about what to assess. There is, for example, evidence about the effects of certain specific program characteristics (e.g., activity-based science) on commonly measured student outcomes (Bredderman, 1983; Shymansky et al., 1983), and one can identify other characteristics that are conceptually or logically related to a wider range of desired science education goals, including science-related experiences that are highly valued in their own right. Looking at the literature through these lenses, three global program characteristics emerge as "ideal" targets for assessment:

- **Access to science knowledge (broadly defined) the extent to which schools provide students with opportunities to learn various domains of knowledge and skills**
- **Press for science achievement a set of conditions related to the expectations schools hold regarding how well students will achieve in science and the degree to which teachers and students act on these expectations**
- **Professional science teaching conditions those conditions that appear to empower teachers and administrators to create science programs in which access is maximized and press for achievement is a dominant feature**

These three sets of school characteristics can help specify the central role of program quality in the educational process and thereby provide a fuller picture of schools in science education. Moreover, though these constructs, on their face, may seem to focus on intangible school climate characteristics, each results from concrete decisions about how to allocate resources (e.g., how much time to devote to science instruction, what kinds of textbooks to buy, what kinds of teacher qualifications to demand and pay for, what kinds of in-service opportunities to offer; what structures to create; and what processes, norms, and relationships to establish at the school). As such, they are alterable characteristics and of interest to educators and policymakers. Thus, assessing these three sets of characteristics should encourage schools to broaden their emphasis beyond raising test scores. And, finally, measuring these program characteristics is likely to help policymakers understand better the conditions under which various science outcomes and experiences accrue. But rather than seeing access, press, and professionalism as being important for their possible direct effects on outcomes, they will be more useful if they are considered enabling conditions that is, to the degree that they exist in schools, they appear to promote (but not to guarantee) high-quality science teaching and learning. This understanding should help inform decisions about what improvement initiatives will be most fruitful, whether undertaken by an individual school or at the district, state, or national level.

The following sections attempt to demonstrate the importance of these characteristics of elementary science programs and suggest how to measure these somewhat intangible constructs.

Access to Science Knowledge

Because what students actually learn at school is influenced by what knowledge and skills they have an opportunity to learn, access to science knowledge can be directly linked to student outcomes. Access is a combined function of school resources, structures, and culture. Basic resources constitute the time, facilities, materials, and staff necessary to bring students in contact with the curriculum (the factual science knowledge, including constructs and principles; the laboratory and science thinking skills; and the general thinking skills to be learned). Of critical importance at the elementary level is time. Generally, the curriculum structure at grades K-6 minimizes the amount of classroom time available for science learning. It also determines the way students are grouped for instruction, generally based on their reading levels rather than the potential contributions they can make to the science learning of the group. A note of caution is in order, however. The current stress on time for science in elementary school has come about because of its virtual absence in most schools. But time for science, at the elementary level particularly, should not be seen as time in competition with time for developing language and communication skills and arithmetical and other quantitative skills, or even social studies, art, or music. All these can and should be taught to some extent in the context of science lessons, and science can and should be part of lessons in these other fields part of the time, while still giving each subject concentrated attention of its own. Thus, the problem of how to count productive time devoted to science in the elementary grades is not a simple one (Raizen and Jones, 1985).

A second important factor is the quality of the curriculum content as embodied in material chosen -- frameworks, textbooks, hands-on exercises and laboratory materials, auxiliary reading (trade books, etc.), audiovisual materials, and availability of computers. Of course, the availability of high-quality curriculum materials does not guarantee

effective use (Raizen, 1987), but their absence severely constrains a science program.

Other organizational structures that enhance access to science learning are programs offering tutoring that provide extra academic support for science learning and extracurricular enrichment activities (e.g., participation in science fairs, field trips, visiting experts, and cooperative programs with museums and universities). Also important are the opportunities the staff have to develop skills for working with culturally diverse groups of students in science and how often schools involve parents in the teaching and learning process.

Assessing the access to science knowledge a school program provides, then, would entail measuring the following more tangible characteristics:

- Instructional time devoted to science
- Classroom assignment practices (ability-grouped or mixed instructional groups) and the curriculum associated with each ability group
- Availability of high-quality instructional materials, laboratories, computers, and equipment, as measured against explicit standards that match curricular goals
- Teachers' qualifications and experience in science
- Use of science specialists or resource teachers
- Availability of academic support programs (tutoring, after-school remediation, etc.)
- Academic enrichment and support (science fairs, field trips, museum programs, schoolwide assemblies)
- Parents' involvement in science instruction or science activities
- Opportunities for staff development in science
- Staff perceptions about the importance of science for all students

Press for Science Achievement

In school programs with a strong press for science achievement, it is clear to both teachers and students that science teaching and learning are taken very seriously and that high achievement in science is expected and valued. Underlying this expectation is a strong belief that all students are capable of learning the important science knowledge and skills schools want to teach (Stevenson, 1986). An atmosphere characterized by high learning expectations is often cited as a key attribute of effective schools (Clark et

al., 1984; Hawley et al., 1985; Purkey and Smith, 1983; Rutter, 1983). Though most research supports the link between expectations and student learning, what helps or hinders students most are the educational structures and processes generated at schools as a result of these expectations. Press is also manifested in how the school's resources are spent and how time and activities are organized.

Press for achievement is gauged by the degree to which administrators, teachers, and students see science teaching and learning as among their most important tasks. When the press is high, students are engaged in a rich and rigorous science curriculum, and they are provided the support they need for success. Science achievement is recognized, highlighted, and rewarded. Noninstructional duties do not interfere with the teachers' primary responsibility to provide good instruction, and science lessons are not interrupted by school routines and nonacademic activities. Administrators spearhead schoolwide policies that create a calm and orderly (not oppressive) atmosphere conducive to science teaching and learning but that recognize that hands-on science is sometimes messy and seemingly disorderly and that deeply engaged students are not necessarily quiet students. Teachers relate to one another as educational professionals in the business of effecting science learning; matters of science curriculum and instruction are part of their collegial work. Teacher evaluation is focused on teachers' skills at engaging children in rich science content and using pedagogically appropriate instructional activities.

Assessing a school program's press for science achievement, then, would entail measuring the following more tangible characteristics:

- Opportunities for schoolwide recognition of science accomplishments
- Curriculum and instructional activities focused on challenging science topics and constructs
- Faculty expectations about students' ability to learn science (e.g., whether all students are capable of learning science)
- Faculty emphasis on science as a subject for elementary school children
- Faculty assignment of science homework
- Instructional leadership in science -- the extent to which a significant person or

group at the school advocates and supports science curriculum and instruction
- The extent to which science teaching and learning is central to teacher evaluation
- The extent to which noninstructional constraints interfere with science activities

Professional Conditions for Science Teaching

Professional conditions for science teaching are demonstrated in the way resources are used, the way programs are developed, and particularly, in the relationships between school administrators and teachers around the science curriculum and instruction.

A professional science teaching climate is a central program characteristic since it comprises the working conditions that are most likely to attract high-quality teachers competent in science and encourage those already in schools to stay (Rosenholtz, 1985). Moreover, the quality of science education that teachers provide in a particular school is enabled or constrained by what the adults and children expect to take place and how they relate to one another. This set of norms reflects basic beliefs, values, expectations, and relationships that shape the school culture. The school culture reflects whether students and teachers are satisfied with their school, whether they believe the school provides a good education, and whether or not students are learning. Together with the resources available and the organizational structures of a school, these climate characteristics influence whether teachers are able and willing to provide "mind-stretching" learning opportunities in science and whether students are willing to take advantage of them. A more extensive discussion of school conditions that enable good science teaching can be found in the Center's companion report on Teachers and Teaching. Here we provide an overview of some of these characteristics and suggest that they, too, should be included in assessments of program quality.

At schools with a high level of professionalism, teachers are committed and energized, permitted to teach science well, and willing to learn to teach better. Staff turnover is likely to be low and stable, and long-range plans can be made and carried out by a cadre of faculty committed to the school (Little, 1982; Rutter, 1983). Professional

conditions cannot be directly linked to science outcomes, but it seems clear that a truly professional staff will work continual on implementing strategies and programs to enhance these outcomes. Because of its importance to teacher commitment, satisfaction, and even if indirectly, teacher effectiveness, the professional climate for science teaching is important to assess. Further, like access and press, professionalism is inextricably tied to educational policies (Darling-Hammond and Hudson, 1986; Darling-Hammond et al., 1983). It is enhanced or inhibited by decisions about resource allocation, decision-making authority, and teacher evaluation, to name only a few.

Assessing a school's professional conditions for science teaching would entail measuring the following more tangible characteristics:

- Teacher salaries
- Teachers' pupil load and class size
- Clerical support staff available for noninstructional tasks
- Teacher time available for professional, nonteaching work
- Time spent on school-based, collegial goal setting; staff-development; program planning; curriculum development; instructional improvement; collaborative research; etc.
- Participation of the staff in schoolwide decision-making
- Staff certainty about their ability to influence and achieve school goals
- Autonomy and flexibility provided to the staff in implementing the science curriculum
- Administrative commitment and involvement in science curriculum and instruction
- Administrative support for professional risk-taking and experimentation

Promise and Limitations

Access to science knowledge, press for science achievement, and professional conditions for science teaching are likely to function synergistically within a science education program. A broad access to knowledge and a press for achievement undoubtedly are most powerful in combination -- when important knowledge and skills are extended to the broadest range of students and a powerful normative force exists that compels and supports teachers' and students' attention to learning. Without a press for achievement, schools providing broad access to knowledge might fall into a pattern of trivializing

science, perhaps by providing a smattering of topics and skills in a smorgasbord of classroom activities. On the other hand, press without broad access might result in schools with elite science programs for only a few students and a vacuum of learning opportunities for the rest.

Ultimately, access and press are unlikely to take hold at schools unless the level of teaching professionalism in science is high. Unless schools have a climate characterized by a belief in the staff's ability to produce high levels of science achievement, academic press and access are unlikely to follow. Conversely, press and access are certain to feed a school staff's sense of professionalism, and they, above much else, nourish professional commitment. This synergy among access, press, and professionalism makes assessing all three sets of characteristics important.

As with outcome measures in science learning, however, current ability to measure these important program characteristics is limited. One possible approach, analogous to the compilation of student profiles and logs of accomplishments in science, is an overall review of the science program by a visiting committee of experts comprising scientists, researchers in science education, science teachers, elementary school principals and teachers, and science-interested parents. Such a review needs to be preceded by careful delineation of the critical components of the science program to be reviewed, such as are suggested in the above lists. The review is more likely to lead to improvements if the school and district staff are actively involved in the review. An example is given in the next section. This approach is feasible at the school or district level. If these reviews are done with care, they could be influential in improving the quality of a school's or district's science program.

At the state or national level, however, collection of information on program features will probably have to be limited to data about the best available proxies for access, press, and professional conditions, though such data could be supplemented with in-depth case studies (Stake and Easley, 1978). This sort of information is likely to spur efforts to improve understanding of science education in elementary schools and, as

better methods are developed, even the capacity to measure its central features. However, such development will occur only if program assessments are accompanied by studies that both analyze the usefulness of current indicators and push the development of a more sophisticated set. In the meantime, a less-than-perfect assessment of program characteristics can provide useful information about the quality of school science programs, help prevent them from emphasizing performance on narrow outcome measures, and provide policymakers and educators with clues about potential problems and promising directions for improvement.

Effective Self-Assessment of the Science Program

As we have described, most external and internal assessments of science focus on student outcomes. Few examine program features. Thus, the characteristics we have suggested for assessment in the previous section are a rather dramatic departure from the traditional focus. Nonetheless, there are some longstanding mechanisms for program assessment (e.g., accreditation processes), and new efforts are now being established.

Many states developing accountability systems have recognized the insufficiency of standardized tests to account fully for a school's or district's performance. Some of these states are designing measures of program characteristics to augment outcome indicators. A few states and districts are pioneering self-assessments of educational programs. The belief that drives self-assessments is that if those actually in the school or district generate and analyze information about their programs, they will use this assessment information (certainly more than they will use external assessment results) for program improvement.

Some groups have developed guidelines for schools embarking on self-assessments. For example, the National Science Teachers Association provides a plan for self-assessments of science programs. The association's plan consists of checklists for principals that cover a wide array of program characteristics. The plan also provides a method of converting the principals' checklist into a matrix that compares current program

characteristics with what they would like to achieve.

Another example comes from the Virginia Department of Education (1986). Their Science Education Program Assessment Model: A Resource Guide focuses on those elements believed to create a more effective learning environment. The Department designed the guide to be used either by external evaluators or for the local school/community assessment teams. The guide includes a data collection and analysis plan, complete with questionnaires for administrators, teachers, students, and parents and a structured classroom observation instrument. It also offers a set of model criteria that schools can compare their results to.

The Weston Public Schools district in Massachusetts has developed a third and somewhat different approach to self-assessment of program quality (Crissman, personal communication, September 1988). The Weston model attempts the simultaneous goals of collecting good data about programs and enhancing communication and trust among the professional staff, scholars, school community, and the public. Each of these groups is represented on Weston's review committees. Their charge is to investigate questions or issues generated by the program staff, administration, and parents. Similar to accreditation processes, the staff compiles background materials for the review committee's use in discussion with the staff and as a guide for observations and interviews. The committee submits its draft report to the school staff and the school committee (school board). The school staff then responds to the draft, and a series of discussions begins in an attempt to reach consensus about the contents of the final report. Because the entire process can take a year or more, the assessment becomes a part of the program itself.

All these models entail considerable confidence that those in and around the school setting can follow a predetermined procedure for collecting and analyzing data about program characteristics, and that the results of such efforts will lead to program improvement. However, much of that confidence rests on the ability of schools to use externally developed guidelines to generate self-evaluation. It also rests on their

willingness to describe and interpret the characteristics of their programs free of bias, values, and opinions, and separate from the struggles among various interest groups that stand to gain or lose from the results of such an assessment.

This confidence may be exaggerated. The nature of a science program is likely to be far more complex, dynamic, and interacting than data from checklists, questionnaires, or structured observations can convey. Respondents and data analysts may have a great deal of trouble capturing and measuring the most important features. And it is hard to imagine that collecting, reporting, and interpreting program data will be free of political influence. Moreover, program improvement based on self-assessment data may fail to bring about intended results, and as with external assessments, they may have contrary, unintended, and unpredictable consequences.

The intent here is not to argue against self-assessment. Rather, it is to suggest that, as schools and districts engage in it, they acknowledge its limitations. As with all such inventions, the contribution of self-assessments to the improvement of science programs will depend on the thoughtfulness with which they are designed and the findings applied. Giving thought to which program factors to assess -- those entailed in access to science knowledge, press for academic achievement, and conditions for teaching, as we suggest, or others -- may well trigger a dialogue about what science programs ought to be and how their goals can best be accomplished. The value of an assessment will also depend on the degree to which those in schools see the assessment process as valid and useful for their own science teaching. The data generated by honest self-assessments that involve science teachers as respected participants can undoubtedly advance the dialogue and lead to effective change.

Self-assessments, at their best, can bring new knowledge to bear, stimulate more thorough discussion and debate, and suggest creative new solutions to the problems of science education in elementary school.

V. IMPROVING ASSESSMENTS IN ELEMENTARY SCIENCE EDUCATION

Improvement Goals

In this chapter, we make recommendations intended to (a) directly assist Ms. Lopez and her many elementary school colleagues through developing and making available good assessment exercises and strategies for their use and (b) improve externally mandated, broad-scale assessments so that they support and encourage the kind of excellent science instruction that Ms. Lopez provides. Before presenting our specific recommendations, we summarize key points made in the preceding chapters and state important improvement goals.

Key Point 1

Assessment can play a critical role in raising awareness among policymakers and the public about the importance of science learning for America's young people and about serious deficiencies in present science learning outcomes. Assessment can also help define the content of that learning. Important constituencies will quickly come to identify the outcomes assessed as those that are important.

Key Point 2

Externally mandated assessments grounded in a full and rich conception of scientific knowledge, skills, and dispositions could communicate to policymakers, the public, and even the education community a bold new vision of science education. Properly constructed and used, such tests can provide sound information about students' knowledge and skills. This information is essential for the formulation and evaluation of educational policies. In these ways, tests are powerful tools that can help improve curriculum and instruction.

Key Point 3

Assessment and curriculum and instruction are interactive. Ideally, instruction and assessment both flow from and inform curriculum goals, and authentic assessment and instruction help shape each other. In reality, the influence of assessment manifests itself in two ways. It is widely recognized that content not assessed is less likely to be taught, but equally or more important, the forms of assessment can come to drive the forms of instruction in undesirable ways.

At the classroom level, students will use teachers' tests to figure out how to study as well as what to study. More globally, the instructional activities will come to resemble testing activities. If narrowly focused tests are used, instruction may also become more narrow. For example, if assessments are used that do not call for extended responses or complex reasoning, the amount of instructional time devoted to these activities may diminish. Thus, the influence of assessment can have negative aspects when assessments are not well matched to curricular goals.

Key Point 4

Interest in science learning outcomes and externally mandated science assessments are increasing together (U.S. Department of Education, 1988). Historically, the power of externally mandated testing to shape curriculum and instruction has been seen in negative terms - as a factor to be minimized. More recently, as states have assumed a more active role in determining curriculum, policymakers have seized on testing as a tool for deliberately shaping curriculum and instruction. Testing what students are expected to learn, the argument goes, will create an incentive for them to be taught what they are to learn. Unfortunately, few of the current assessments reflect modern understandings of the range of important science learning outcomes.

Multiple-choice, paper-and-pencil exercises predominate, and the focus is far more on recalling facts than on understanding important constructs and principles of science and acquiring the skills integral to science.

These points undergird the need for improving assessments of science learning so that they will support and guide exemplary science education. We identify three improvement goals:

> **Improvement Goal 1.** Making classroom assessment an integral part of ongoing instruction.

Teachers should be given education and experience both in selecting a variety of short-term and long-term assessment strategies and in using them for different instructional purposes -- determining what science knowledge children bring to a lesson, observing what prerequisite science skills they have, tracking their learning progress and the effectiveness of the science instruction, organizing productive working groups, and making judgments about individual and group attainment in science over time.

Teachers should be trained to evaluate their own tests and assessment strategies as well as externally mandated tests; principals and school or district science specialists also need to be able to evaluate the quality of tests and their correspondence to the school's or district's goals and objectives in science education.

> **Improvement Goal 2.** Development of externally mandated assessments as well as classroom tests that conform closely to the characteristics of good science curricula and instruction, as enunciated in this report and in the Center's two companion reports on curriculum and instruction and teachers and teaching.

Assessments should (1) provide greater opportunities for children to interact with stimulus materials, (2) attend to understandings of constructs and principles as well as factual knowledge, (3) probe approaches to problem solving as well as outcomes, (4) be explicitly integrated with the curriculum and with instruction, (5) incorporate hands-on activities wherever feasible, and (6) be structured around group as well as individual activities. Development and validation of such tests will require close collaboration among content-matter specialists, experts in science curriculum, persons knowledgeable

about the realities of the classroom, and psychologists and psychometricians.

> **Improvement Goal 3.** Ensuring correspondence among assessments conducted at different levels, that is, creation of assessments that are likely both to encourage better science programs at the local level and to inform policies at a more global (e.g., state) level so that such policies will be effective in supporting local improvement efforts.

State-level policymakers are necessarily concerned with a different set of alterable variables than are educators in districts and schools, but both sets must be grounded in a common, coherent, and comprehensive conception of learning outcomes. Assessments of the kind envisioned by our panel can serve audiences at both of these levels. For example, an assessment that reveals deficiencies in students' use of science tools and thinking skills needed in science may trigger greater emphasis on those skills in classrooms and also highlight the need for states to provide technical assistance and resources or to address hands-on science teaching in a state's teacher education programs.

> **Improvement Goal 4.** Attention to careful and informative analysis, reporting, and dissemination of assessment results.

Reference standards for test scores should be based on the development of science understanding and science-based skills, not on national norms designed to rank-order students. Oversimplified summaries of test results must be avoided. Different incentive structures are created by reporting at different levels of aggregations. Information must be provided in a form and at the level at which it can best guide improvement at the local level. At the same time, simplistic rankings of schools, especially when reinforced by rewards or sanctions, may quickly erode the validity of the assessment, leading to efforts to improve test scores without bringing concomitant improvements in the student knowledge, skills, and dispositions that those scores were intended to represent.

A Starting Point

What needs to be done to change the current unsatisfactory state of assessment of science learning in elementary school, to move toward the kinds of assessments described in the preceding chapters and represented by the three improvement goals? A first question is where to start. Should efforts to improve focus on the classroom level where testing has the greatest effect on children's future learning and engagement with science? Should they focus on broad-scale assessments conducted at the national and state levels since these catch the attention of the media and the public? Or should they focus on the tests specific to a state's or district's curriculum since real sanctions and incentives are more and more commonly being attached to assessment results at these levels?

In the panel's view, improvement of assessment must proceed at all three levels simultaneously. An interesting prototype that might serve as starting point and provide guidance is Great Britain's new assessment design, even though this design envisages a national science curriculum -- an unlikely prospect for the U.S. Indeed, in a country as large and with as diverse a set of educational systems as the United States, Great Britain's tightly integrated approach between the teacher's need to assess for instructional purposes and the national, state, and local needs for monitoring, accountability, and information to devise better policies is probably not possible. What the British approach does imply, however, is that improvement must be fostered simultaneously at the classroom and the broad-scale assessment levels -- just as our panel suggests.

The recent report by a task group on assessment and testing convened by the Department of Education and Science and the Welsh Office (1987) outlines the integrated assessment system recommended for use in Great Britain at all levels. The system is based on a combination of moderated teachers' ratings and standardized assessment tasks. The teacher ratings would themselves be based on the many sources of information that a teacher like Ms. Lopez uses to assess a student's progress,

including general impressions, marking coursework, marking assignments, student self-assessment rating scales, checklists, practical tests (of hands-on performance), and written tests. The scales being used by teachers to rate students on the basis of these various sources of information would be made generally comparable through a process of bringing individual judgments into line with general standards (moderation). The standardized assessment tasks would include written test responses, practical (hands-on) tasks, and observations covering the several goals of science education. Standard assessments would take advantage of several presentation modes: The question could be delivered orally, in written form, pictorially, through video or computer, and through practical demonstration; the expected method of student work could be mental only, written, practical, or oral; and the response mode might vary from multiple-choice questions, writing a short prescribed response, open-ended writing, oral response, practical procedure being observed, practical outcome, or product or computer input.

In the proposed British system, assessment results derived from the same sources would be used for the teachers' classroom purposes and aggregated for reporting at the school, district, or national level as a way of assessing student learning and evaluating the quality of science programs. Many problems of validity and reliability of the various assessment modes remain to be addressed, but this thoughtful and comprehensive design warrants close attention.

A Systemic Approach

Several functions must be in place to create an integrated assessment system responsive to educational goals at each level of the system. Foremost among these are:

1. The development of imaginative, creative assessment exercises that will probe performance and higher-order thinking skills and the application of these skills to new situations. This will be a time-consuming and costly effort, particularly if some of the exercises are to take advantage of the availability of computers and other information technology or if, in the case of hands-on performance needing observation or interview protocols that require interpretation, scoring rubrics and standards leading to reliable results need to be developed.

2. The capability to design and make widely available assessments using good individual exercises and other assessment strategies to fit particular purposes and contexts.

3. A mechanism to ensure the high quality of individual assessment problems and exercises, other assessment components, and whole assessments provided for use at the classroom level or for larger-scale assessments conducted for purposes of monitoring and formulating policy.

4. Availability of assistance to teachers and district and state personnel in selecting and carrying out appropriate assessments and in analyzing, interpreting, and reporting results.

5. An ongoing program of research to increase correspondence between assessments and changing learning goals in science education and to ensure that both instruction and assessment utilize advances in knowledge (e.g., how children learn) and in technology (e.g., using computers for tracking problem-solving strategies).

Some of these functions may best be carried out at the national level, some need to be decentralized so as to work effectively at the district and school level. Generally, research and development functions entail high risk -- much will have to be discarded as development of assessment exercises and strategies proceeds -- and require investment in the best talent available (Committee on Research in Mathematics, Science, and Technology Education, 1985; 1987). This implies that functions 1 and 5 need to receive attention at the national level. Capacity for functions 2 and 3 -- designing appropriate assessments and ensuring quality control -- needs to be built at all levels. In education, building capacity often has been a combined responsibility; for example, under Title II of the Education for Economic Security Act (EESA), funds are made available to states by the U.S. Department of Education to assist local districts that have developed good plans for improving science and mathematics education. On the other hand, function 4 -- staff development and assistance as well as dissemination -- needs to build on local and intermediary structures already in place for these purposes but not equipped to deal effectively either with science education or with assessment. In addition, national and state efforts must be accompanied by local experimentation to develop improved classroom-level assessment.

National Functions

To say that a set of important activities needs to be initiated, supported, and maintained at the national level does not necessarily mean a centralized set of activities. The distinction is well understood in basic research, where nationally funded support mechanisms have been created for example, through the National Science Foundation (NSF), the National Institutes of Health, and the Office of Naval Research that allow many institutions and talented individuals to participate. Development efforts resemble research when products cannot be prespecified, as is the case for developing assessment exercises and strategies. An analogy is the set of eight projects being supported at different institutions by NSF to develop materials for elementary science education; at the same time, private foundations are also supporting curriculum development for elementary science. Below, we take up in greater detail our recommendations for research and development to provide the basic building blocks necessary for improving assessment of science learning at all levels -- valid assessment exercises, alternative assessment strategies, and assistance with quality control and appropriate application.

> **Recommendation for Research:** We recommend that cognizant federal agencies and private foundations undertake a program of research designed to improve the foundations underlying science assessment in elementary schools. The research should be directed toward two goals: (1) increasing educators' understanding of what should be assessed and (2) improving the methods for collecting information about students' science learning.

Specifically, we suggest four research areas, two addressing the first goal and two addressing the second goal. We believe that federal agencies concerned with science and with education, particularly NSF and the Department of Education, should undertake support of these programs at a level not below $5 million a year and preferably $10 million a year.

Finding Out "What Matters" in Science Education. Research programs that will provide insight into what should be assessed -- what matters in science education -- need to

address two broad and extremely complex areas. The first relates to expanding what is known about cognition in science. There has been far more research on the cognitive processes involved in language acquisition and reading comprehension than on those involved in learning the multifaceted aspects of science. Until assessments parallel that kind of learning, they will continue to be restricted to measuring rote memorization and fail to measure students' acquisition of the most important aspects of science learning goals.

The second area that must be addressed in establishing what matters in science education relates to identifying those aspects of science knowledge and skills that have the greatest benefit for maximizing human potential. For example, what science learning is most likely to benefit individuals on the job as well as in other daily life situations? What science knowledge and skills shared by citizens are most likely to contribute to the betterment of society as a whole?

Research Area 1: Theory and empirical base. In addition to the outcomes or products of science learning traditionally considered important, assessment of what matters must include the intellectual skills required to apply what one knows about science to learning more about science and to solving academic and real-world problems. By increasing their understanding of how students learn science, educators could dramatically improve instructional effectiveness and thereby improve the quality of assessment in the service of instruction as well. Moreover, if such abilities as solving academic and real-world problems and conducting inquiries to learn more about science are major goals of science instruction, a far better understanding is needed of the interrelationships among "knowing" science, being familiar with the methods for conducting science, and understanding the component skills of problem solving. Why is it that some students achieve the structure and coherence of skills and science knowledge essential to science competence whereas others do not?

Much more should also be learned about the relationship between the nature of one's science knowledge base and its successful application to problem solving and learning.

First, the characteristics of the science knowledge base that facilitate learning, including its structure and how a well-structured knowledge base is developed, need to be identified. Second, the relationship needs to be defined between the characteristics of the knowledge base and the intellectual (thinking) skills relevant to science. For example, do some individuals have knowledge bases that are structured so they are easier to access than others? Do certain qualities of an individual's knowledge base facilitate useful connections?

Research Area 2: Relationships between science learning and effectiveness in life beyond school. Why should students study science in school? Although they live in a society permeated with and dependent on science and technology, recent studies (International Association for the Evaluation of Educational Achievement, 1988; Mullis and Jenkins, 1988) show that students know little about science and technology and perceive that the science they study in the classroom has little relevance to their lives. It has been posited by numerous study panels that increased understanding of science is vital both to individual careers and to the economic health of the nation (National Governors' Association, 1987a, 1987b; Task Force on Education for Economic Growth, 1983). Though studies have estimated that a high percentage of all jobs will require some direct understanding of science and technology (Education Commission of the States, 1982, but see Levin and Rumberger, 1983, for counterarguments), there has been little research on the connection between the more intangible learnings gained from formal science experiences and success in the workplace.

Looking toward the year 2000, the fastest-growing occupations will require employees to have much higher reasoning capabilities than do current occupations (Hudson Institute, 1987). Incoming college freshmen are not generally considered at-risk, but neither are they ready for the work force, at least as diagnosed in one study (National Alliance of Business, 1987:5). The question that must be raised, then, is to what extent additional training in science contributes to higher reasoning capabilities and to work force readiness. For example, will students who have planned and conducted many science experiments actually be able to perform better in the work force? Do they ask for

evidence and analyze data more effectively than employees who have had little or no experience with the methods of science?

Such lines of research should not be confined to the impact of exposure to science on success in the workplace but should be extended to study the impact of science learning on the quality of life for individuals in nonwork settings as well as its impact on society as a whole. For example, it would be useful to know whether individuals with science training feel a greater sense of confidence or empowerment in social situations, in dealing with family situations, or in addressing their own health needs than individuals with limited exposure to science. In addition, one must ask whether society would be better off if more people understood the multiple hazards to the environment, issues of national defense, research to cure diseases and increase longevity of life, and the relationship between scientific discovery and global competitiveness. Thomas Jefferson felt that the survival of democracy depended on an enlightened citizenry. In other words, if people do not understand the issues, they cannot make intelligent decisions, and the process of self-government or "government by the people" will not work effectively. What is the role of science education in creating an informed public?

Assessing "What Matters" in the Best Possible Way. To research ways of improving measures of science learning entails finding more effective and efficient methods for developing such measures (Committee on Research in Mathematics, Science, and Technology Education, 1985; 1987). This requires that considerable time, energy, and resources be devoted both to investigating students' reactions to problems, and to interviewing students about how they interpreted assessment questions or tasks and why they responded the way they did. These investigative and interactive procedures should be directed equally toward improving the validity and reliability of the assessment instruments. Only after iterating these procedures on a small scale should resources be dedicated to conducting larger pilot tests.

Research Area 3: Improving the validity of science assessment measures. If an established empirical understanding of the cognitive processes underlying science

learning and problem solving existed, it would be crucial to improve the link between these cognitive processes and the assessment questions presented to students. For example, if a particular cognitive structure makes students' science knowledge bases more conducive to problem solving, one would want to investigate if and to what extent students' knowledge bases were so structured. In addition, learning more about the influence of the knowledge base and thinking skills on the strategies students select to perform assessment tasks is critical to the assessment of science-relevant intellectual skills. The "right" answer for the wrong reason may indicate far less learning than a "wrong" answer obtained using more effective or advanced cognitive strategies. Verbal protocols or computerized records of students' efforts in problem solving may prove fruitful avenues for further research in these areas.

In addition to better measures of cognitive processes and a well-structured knowledge base, more effort should be devoted to designing better measures of proficiency in science-related laboratory and intellectual skills, including the appropriate use of scientific equipment and the principles underlying the conduct of investigations. Although hands-on assessment techniques have long been integral to assessing science in Great Britain and have been explored in the United States (Blumberg et al., 1988; Connecticut State Department of Education, 1986) and in other countries, some hands-on tasks seem to isolate skills (much as decoding in reading) by asking students to measure or observe particular phenomena out of context.

Research Area 4: Improving the reliability of science assessment measures. As the procedures for assessing science learning become more rigorous and complex, the measures will become increasingly sensitive to the contexts in which they are collected and the ways in which they are interpreted. For example, research should be conducted about the effect of inclination on measures of science learning. There is a vital distinction between having the ability to do something and having the desire to display that ability on demand. It is possible that some students may be too shy to explain their thinking to interviewers, and others may simply decide against engaging in difficult assessment tasks if there is little apparent reason to do so.

Finally, as the assessment measures become more varied and sophisticated and the behaviors assessed more subjective and complicated, research will need to establish reliable methods for evaluating observed or recorded student performance. As the measures correspond more and more to valued outcomes for science education, those interpreting students' responses to tasks will have to develop ways of decreasing ambiguities in interpretation to the point of routinized agreement. Further, the results of these evaluations, as well as the methods used to obtain them, will need to be articulated in ways clearly understood by general education audiences.

Disseminating Research Results. The complete set of findings from this research agenda must be disseminated to have any effect. If evidence for what matters in science learning is weak and the crux of science education is too difficult to articulate, parents and legislators will not care whether students know and can do science. And unless the new theory-based type of assessment we advocate and the methods used to implement it can be communicated to teachers, administrators, and textbook publishers in clear and and compelling ways, it will never be widely recognized, and the improved assessment procedures it yields will seldom be used.

> **Recommendation for Development:** We recommend that cognizant federal agencies, states, and test developers undertake the development of assessment exercises and assessment strategies designed to probe the various understandings, competencies, and dispositions that make up the goals of elementary science education. The exercises need to address performance competencies as well as paper-and-pencil responses, open-ended tasks and questions, and evaluation of learning over time and in groups. To accompany the exercises, careful protocols for interpreting observed behaviors and responses must be developed. Assessment strategies must be devised that address information needs at different levels of aggregation and that incorporate informal as well as formal means, as appropriate.

The development of improved assessment exercises and strategies cannot await the results of the research program. First, the need for better assessments is too great to delay any longer, and second, the two sets of activities should be supported concurrently so that they can inform each other. They must also proceed in concert with curriculum development in elementary science so that assessment corresponds to the most innovative and effective curricula, and both curriculum development and assessment need to draw on the ongoing research. Four development areas deserve special emphasis: (1) creating performance tasks for individuals and groups and accompanying protocols for rating performance; (2) creating assessment exercises, including some open-ended situations, that provide sufficient time for sustained work, again accompanied by effective rating protocols; (3) developing methods for teachers that would allow them to document their students' progress in a systematic fashion; and (4) exploiting the computer's potential for tracking students' thinking as they address science problems.

Development Area 1: Performance tasks and rating protocols. Experimental work has been supported in this area to the extent that several states are examining the possibility of including a few performance tasks in their state science assessments. Specifically, the National Assessment of Educational Progress (NAEP) conducted a pilot project (Blumberg et al., 1986) in which adaptations of the exercises developed in Great Britain by the Assessment of Performance Unit (1984-1985) were tried out with children in U.S. schools. The experimental tasks included group activities administered to whole classes and asked for open-ended, paper and pencil responses to problems posed in various ways. There were also station activities where hands-on tasks required students to use equipment or materials to investigate relationships and then answer open-ended questions based on their findings. These tasks were administered to small groups of students, with the students rotating from activity to activity, some computer-administered. In addition, complete experiments were administered to individual students, with the administrator posing questions, explaining the equipment, and using a checklist to report how students used the equipment to conduct their experiments. After students had completed their investigations, they discussed their findings with the administrator. NAEP concluded after this experiment that "although managing

equipment and training administrators requires ingenuity and painstaking effort, conducting hands-on assessment is feasible and extremely worthwhile." A summary description of this effort is available from NAEP (1987).

Several states, including California, Connecticut, Illinois, Massachusetts, and New York, are attempting to develop assessment exercises and strategies that will come closer than can paper-and-pencil tests to probing some of the performance competencies included in their curricular goals for elementary science. In Connecticut, the plan is to work toward a comprehensive assessment strategy that includes sustained assessment tasks that integrate knowledge and understanding with skills and dispositions. In connection with its second state science assessment in 1984-1985, Connecticut conducted a pilot administration of practical tasks involving 900 students. The additional cost per student was $6.66 for administering the items by trained external test administrators and for developing the scoring rubrics (Baron, 1988). Thus, performance tasks appear quite feasible in large-scale assessments as well as for classroom use, provided an adequate number of good tasks are available. This is not now so. Even the British exercises mentioned above, which have been under development for some time, exhibit some problems, particularly concerning their psychometric properties (personal communication, Richard Shavelson, December 23, 1988). Thorough testing of performance tasks is expensive, but necessary (Pine, 1988), and needs adequate funding.

Development Area 2: Assessment exercises involving open-ended situations and sustained work. Almost all current test items, including most performance tasks, prespecify the problem to be solved and set parameters for the solution(s). Obviously, this does not mirror the way real problems in science present themselves; it does not allow students to demonstrate what they might have learned about formulating a researchable problem from a messy question devising alternative research approaches, testing these out, and coming up with an approach that promises to yield useful information. The opportunity to exhibit the integration of knowledge and skills needed to address a science-related question where the problem, let alone the solution approach, is not self-evident also requires that adequate time be allowed the student.

We believe that the development of how to stage such open-ended situations for purposes of instruction as well as assessment would greatly assist the classroom teacher and support important goals in science education. Development would need to include adequate descriptive material on how to set up the assessment situation, how to guide students in their approach through appropriate coaching (when necessary), and how to evaluate their progress and performance. Until a serious effort is undertaken to develop and field-test these sorts of assessment exercises, it is hard to predict whether they will be feasible in large-scale assessments. However, their classroom use alone would justify the investment for the message they would send about the nature of science learning and the model they would provide for good science instruction.

Development Area 3: Documentation of student progress. One of the most regrettable aspects of short-answer tests given at one point in the school year is that they cannot reflect a student's long-term development of a science knowledge base and science-related competencies; all these tests can do is to provide a snapshot of student achievement. One solution is to give the same test at the beginning and end of instruction (pre- and post-testing), as do some of the countries participating in the IEA international assessments, so as to track the effects of instruction. This may show progress on specific items (increasing the temptation to teach them in the case of high-stakes tests). But even if the items have ecological validity (that is, mirror important science learning goals), the retest is likely not to reflect the breadth of knowledge and skills acquired in the interim in a good science classroom such as Ms. Lopez's.

During a recent conference on science assessment in elementary school (Lesley College, Mass., November 4-6, 1988), the need to develop more systematic approaches to documentation of student progress was highlighted by Edward Chittenden of Educational Testing Service. Specifically, teachers need ways to document growth in students' science thinking and in what students are able to do, growth that is not captured by records of their written work and their test scores but by their discussions (Chittenden, 1988), their behaviors as they tackle laboratory tasks and science problems, their questioning, and their monitoring of their own and their peers' work. None of these

important science-related behaviors leave recordable tracks unless systematic efforts are made to document them. Prototypes for developing such documentation need to be created based on research evidence of what behaviors it is important to track in order to assess progress in science learning. When reliable documentation methods have been created, related training materials will have to be developed for use in preservice and in-service education to ensure that teachers use these methods appropriately. Again, a triple benefit would accrue: a message about what matters in science learning, good models for instruction, and much enriched information for assessing student learning.

Development Area 4: Computer-based assessment. Computers have the unique capability of recording people at work on all sorts of mental tasks. Because they log every response and can, if so programmed, adjust the task and provide prompts (coaching) responsive to an individual's step-by-step performance, computers are a potentially powerful assessment tool. So far, both research and development on computer use in science learning -- leaving aside recording of laboratory data and computational uses -- have focused on creating instructional modules: simulations of physical phenomena and sites inaccessible to the classroom, microworlds that model idealized or simplified environments, intelligent tutoring systems that employ coaching paradigms to teach specific knowledge and skills. (For a listing of examples, see a recent report by the Office of Technology Assessment, 1988.) As student access to computers continues to increase (Becker, 1986), the possibility of incorporating such modules into science instruction becomes real. Adaptation for use in assessments seems entirely feasible, for use by the classroom teacher and for larger-scale testing. Although verbal protocols give some insights into children's thinking in science (see, for example, Chittenden, 1988; Driver et al., 1985), the computer can provide much more extensive records uncolored by a human observer's interpretation though the repertoire of student responses that a computer can accept and react to is likely to be more constrained. Computer records of student work, if appropriate means for analysis are developed, could serve as an important database for evaluating students' progress in developing scientific thinking and reasoning skills. The records would also provide a superb resource for further research on how individuals structure science knowledge and

bring it to bear on science-related problems.

Quality Control and Dissemination. As with research, it is critically important that the results of the efforts to develop improved assessment exercises and strategies be disseminated, and disseminated appropriately. Appropriate dissemination entails not only making materials widely available but also evaluating their quality before recommending their use and advising on effective and feasible assessment strategies for given contexts and levels of aggregation.

> **Recommendation for Quality Control and National Dissemination:** We recommend establishment of a center, or network of centers, to collect promising examples of innovative exercises and strategies for assessing student progress in science learning, evaluating their quality and feasibility, and making them available to agencies designing large-scale assessments as well as to intermediaries assisting schools and teachers to devise improved teacher-controlled assessments.

The proposed center, or centers, could be based on existing centers with a related mission (for example, the National Science Resources Center) and be located at universities, as adjuncts to state departments of education, or in private research institutions. We see the center(s) serving as a centralized set of resources and clearing houses allowing people charged with assessing science learning to survey and obtain the very best available assessment materials as well as guidance on appropriate assessment strategies. Besides the collection of assessment exercises, quality control, and dissemination, the functions to be carried out would include designing appropriate combinations of exercises for particular purposes and contexts. This national-level resource, whether one center or several, is seen as being able to relate to many intermediate agencies that exist specifically to serve local needs but not directly to teachers and schools.

Our three recommendations for research, development, and dissemination and quality control mirror a recommendation made by the National Academy of Sciences (Murnane and Raizen, 1988:65):

> that a research and development center be established to provide for the efficient production, evaluation, and distribution of assessment materials for use as indicators of student learning at district, state, and national levels and for use by teachers in instruction and assessment.

We believe, however, that talented researchers and developers interested in improving the assessment of science learning are to be found in many locations and that programs to fund this difficult but important work should encourage their widest possible participation, with due regard to the critical mass of resources needed for any one project or activity. On the other hand, quality control and dissemination cannot be left to the vagaries of voluntary involvement. Because they are generally considered less enticing as an intellectual activity than research and development, they are often neglected. Deliberate investments must be made and means designed for ensuring that the products of research and development are effectively used to improve the assessments used by teachers and the assessments designed for broader monitoring and policy purposes.

Our panel urges that the National Science Foundation, the U.S. Department of Education, and private foundations establish programs of research and development in the assessment of science learning and consider how a quality control and clearinghouse function might best be established to ensure use of best available exercises and strategies so as to improve current assessment practices. We estimate that an investment will be needed of at least $5 million per year over the next five years for research, an additional amount of at least $5 million per year for development, and an initial yearly investment of $1 million for the quality control and clearinghouse function, the latter to grow as more assessment exercises and strategies are developed.

Decentralized Functions

It is our view that this type of national effort must be accompanied by parallel local efforts. At this level, however, improvement strategies cannot concentrate solely on assessment. The contrast with national-level efforts is striking: Because there is now recognition of the need for improved science curricula and teacher development and considerable investment in these areas through NSF, EESA Title II, and private foundations, it is important -- without diminution of the already ongoing programs -- to direct attention to the assessment area, which has been severely neglected and which so critically influences the other two. At the local level, however, means for having direct impact on science learning in most of the nation's elementary classrooms are sadly missing in all areas -- curriculum, teaching, and assessment -- and all three are in dire need of improvement. Hence, we see the need for mechanisms that will serve schools and teachers to improve science education in all respects, with better assessment practices as an important concomitant.

> **Recommendation for Local Dissemination:** We recommend a dissemination system for science education that will put in the hands of teachers the very best science curricula currently available, assist them in designing and using appropriate assessment strategies, and provide opportunities and materials for needed staff development.

A goal of our Center is to synthesize research and exemplary materials that illustrate the many dimensions of effective science education and make that information accessible to educators and noneducators alike. However, the Center's dissemination function is largely limited to production of print materials and some interactions with professional organizations in order to devise cooperative ways of reaching all the audiences interested in science education. Yet, print alone will not suffice, as anyone knows who has studied and understands how improvements spread and are institutionalized in education systems (Berman and McLaughlin, 1975-78; Havelock and Lingwood, 1973; Human Interaction Research Institute, 1976; Rogers, 1962; Yin et al., 1976).

A Dissemination System. The magnitude of the task of reaching all those with some responsibility for improving science education calls for a distinct and unique dissemination system. The goal of such a system is ultimately to change the science instruction received by children. To achieve this goal, many people need to be reached, ranging from federal agencies and other national bodies, to, higher education and school district organizations, to classroom teachers.

The functions of an effective dissemination system are:

- Packaging of information and materials, recommendations, and models of science teaching and learning that emanate from research, development, and special groups created to address problems in science education, for example, drawing on the new curriculum development work, the National Science Resources Center (1988), the exemplary programs identified by the National Science Teachers Association, and the work of the proposed resource and clearinghouse in assessment

- Formulation of the most appropriate delivery strategies, for example, training for teachers or administrators or a combination, training for other school or district staff, awareness sessions for parents and school boards, manuals for teachers and principals, policy briefs for school superintendents and for local and state legislators

- Identification and/or development of delivery systems, for example, existing or needed organizations, agencies, networks, and interaction with them or among them

- Coordination and ongoing support of delivery, with quality control that ensures soundness of and equitable access to information, materials, and services

- Provision of channels for informing policymakers, researchers, developers, and others concerned with improving science education of needs in the classroom as experienced by teachers and local administrators

We suggest below one possible design for a dissemination system responsive to these functions. This dissemination system design has several features. First and foremost, it

is based on research that shows that effective dissemination promotes meaningful changes in practice when it provides a sufficiently high level of assistance in an environment where support and clear expectations create pressure for change (Crandall & Loucks, 1983). A dissemination system must develop the capability of people at several levels through staff development and ongoing support while it works to create a context in schools, districts, and states where there are incentives, resources, and clear direction.

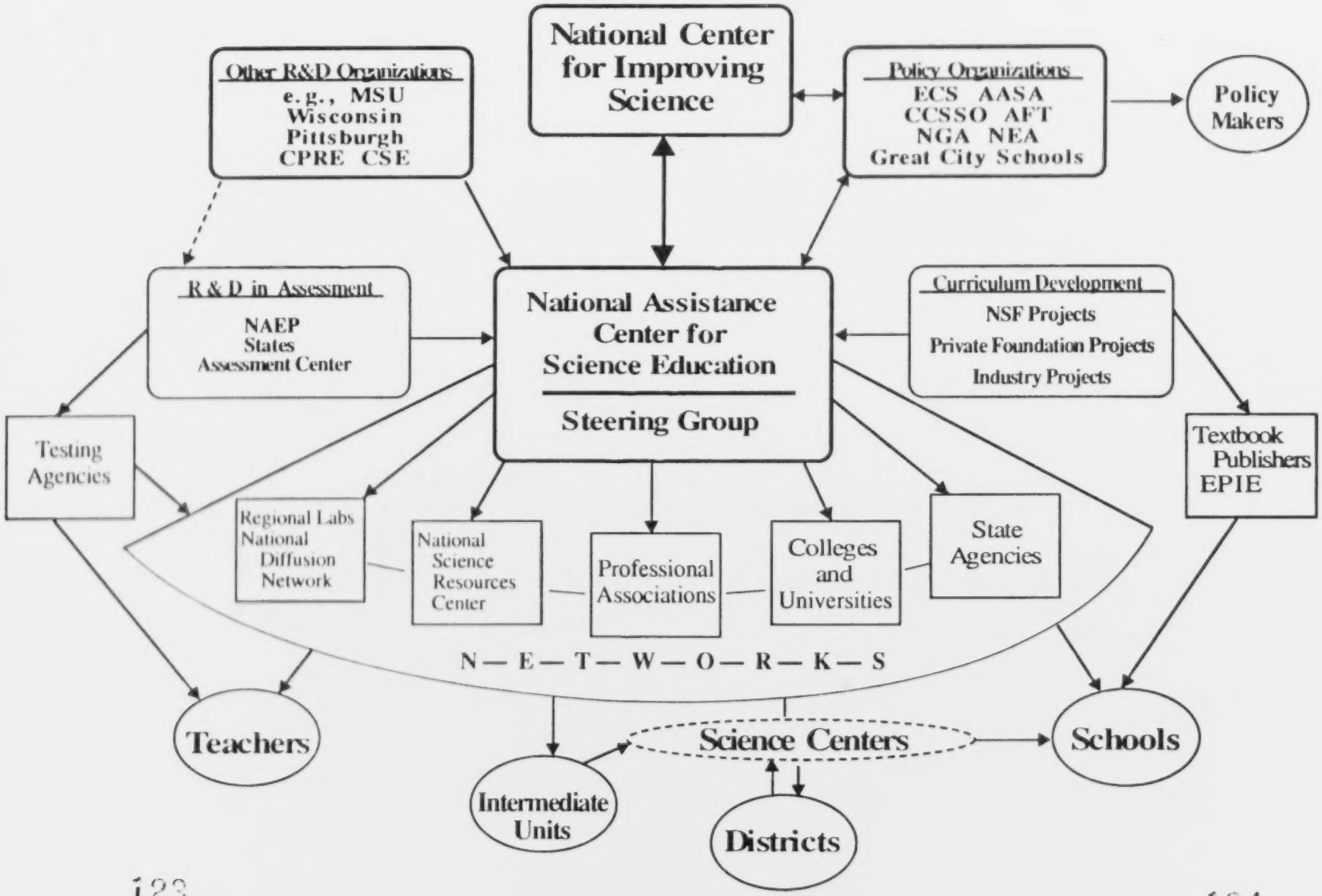
A second feature of any good dissemination system design should be that no part of it is meant to displace or replicate the work already being done by others. Instead, it should enhance their work and piggy-back on their efforts. Related to this is the notion that every function of the dissemination system listed above should be addressed through multiple channels and should serve multiple constituencies. Although this may lead to complexities in understanding the system and difficulties in drawing clean organizational charts, it ensures access. (Because such a multiple system may also lead to dilution of quality, we have suggested that quality control be a centralized function in the case of assessment exercises and strategies.)

Finally, a system designed in this way requires coordination and ongoing support in the form of infusion of new materials, ideas, and strategies deriving from all levels of the educational system, as well as problem-solving assistance. This coordination function also entails mechanisms to ensure the quality of the system's work, including equal access of all populations to the services being provided, with regular assessment of operations and impact.

Figure 2 illustrates the structure of our design for a dissemination system. At its hub is a National Assistance Center for Science Education. This center (or configuration of centers) could be an expanded version of the assessment quality control and clearinghouse center(s) recommended above but also encompass curriculum and instruction, teacher development and enhancement, and improvement of the school context for science learning. Alternatively, the Assistance Center(s) would work closely

Figure 2

Structure of a Dissemination System



with the Assessment Center(s), as it would with curriculum and teacher education groups. The Assistance Center(s) would have primary responsibility for seeing that the functions of the dissemination system described above are carried our effectively, with the framework of its work coming from the National Center for Improving Science Education. Assistance Center staff would solicit and evaluate input from other research and development organizations as well as individuals and special study groups. They would work directly with policy organizations to disseminate to policy audiences, with particular emphasis on the kinds of policies that would provide the direction and support needed to spark improvements in science education.

The system for reaching the practice community is somewhat more complex because of the "multiple access" design. A steering group might represent the primary service providers -- the regional laboratories, professional associations, colleges and universities, state agencies and their intermediaries, and science teacher centers that directly serve local school districts. The function of this steering group would be to advise on packaging of exemplary materials that is effective for the multiple audiences, help formulate appropriate delivery strategies, and identify and link to existing delivery structures. Assistance Center staff then would work with both individual organizations and clusters of organizations identified by the group, helping them incorporate new materials and strategies to better meet the needs and broaden the base of their constituents. Center staff also would solicit input on special needs and on promising practices emanating from the classroom to feed back to research, development, and policy groups.

This general scheme needs elaboration in the case of assessment of science learning. The multiple agencies already in place to assist teachers and local educators would be expected to work closely with them in improving classroom assessment carried out for instructional purposes. They would be expected to act as effective intermediaries between the Assessment (and/or Assistance) Center's bank of assessment exercises and assessment strategies and the needs of particular schools and classroom teachers. They would serve the in-service and staff development needs at the local level to ensure that

teachers possess the requisite skills to select and administer appropriate exercises and a variety of assessment techniques to probe the full range of science education objectives, and that they are able to interpret results accurately. In addition, these service agencies should have the resources and expertise to bring students to centralized locations in order to have them work on hands-on tasks, computer simulations, and other sorts of exercises not practicable in a particular school or classroom.

Unfortunately, except for professional organizations and college science faculty, few of the existing channels and agencies designed to assist teachers and local educators have expertise in science; fewer still know much about alternative assessment approaches. One obvious set of resources to be built on are the existing science materials and teacher centers, where assessment components could be built in with relative ease. Such science-based intermediary institutions have a record of maintaining excellent science programs (Penick, 1983; National Sciences Resources Center, 1986), and they operate successfully set up in large or small school districts, rural or inner city (Anchorage, Alaska; Mesa, Arizona; Schaumburg, Illinois; Fairfax County, Virginia; Seattle, Washington; Milwaukee, Wisconsin), or serving a whole region within a state (Spencerport, New York; Portland, Oregon). But unfortunately, most school districts do not have that sort of science program support available; therefore, effort and resources will have to be invested to build capacity for both science education and assessment in other existing service agencies and institutions.

The dissemination scheme we suggest is not the only feasible one, nor possibly even optimal. Our purpose here is to outline some necessary characteristics of an effective dissemination system, based on research and experience with the regional laboratories; the National Diffusion Network; assistance agencies within states, including science centers; university extension services, and such independent agencies as the Educational Products Information Exchange (EPIE). The ultimate aim is to utilize the resources being developed at the national level to enable teachers, schools, and districts to select, design, and appropriately use assessment exercises and strategies that are consonant with their curricula and probe across all the science learning outcomes they value. An

important concomitant is that the constituent service agencies and organizations be actively involved in the design of needed teacher preparation and staff development to foster use of innovative assessment materials at the classroom level.

> **Recommendation to Design a Dissemination System:** We recommend that the National Science Foundation and the U.S. Department of Education establish a study panel to identify how current dissemination resources designed to improve education need to be enhanced and built on to provide effective services to teachers and local school administrators in improving science curricula and instruction, assessment of science learning, and staff development to build science teaching competencies.

The study panel should complete its work in 18 months. Because dissemination is always more costly than research and development, often by a factor of 10, we anticipate that investment for dissemination and assistance services eventually should be budgeted at a minimum of $100-$200 million a year (e.g., through focused use of monies available through Title II of the EESA). Most of this investment should be targeted for service to schools serving at-risk populations.

A Special Project

It will take several years to develop an effective dissemination, logistics support, and staff development system based on research and exemplary practice in science education. Meanwhile, some immediate steps could be taken to improve assessment of science learning in the classroom. Specifically, evidence indicates that an important influence on the quality of science education is the textbook and the text-embedded quizzes and test exercises. Surveys of the practices of elementary school teachers have shown that they use these quizzes and problem sets fairly extensively to assess their students' achievement. For this reason, we suggest that a special project be funded to address the quality of these materials.

> **Recommendation for a Special Study:** We recommend that a systematic study of the quality of curriculum-embedded science tests be undertaken, and that the results be used as the basis for a conference with textbook publishers and state

and local assessment experts to encourage improvement of these text-related assessment materials.

The conference's aim would be to encourage textbook publishers to develop and include in their texts problem sets and test questions more consonant with the range of curricular goals that states and districts have enunciated for science education rather than concentrating on memorization and rote problem solving. This could provide immediate help to teachers in evaluating their students' science learning and would move assessment in the service of instruction forward while more comprehensive approaches are being developed. The estimated cost for the study and conference is $200,000 over 18 months.

Local Initiatives

Thus far, we have addressed the four improvement goals set out at the beginning of this chapter through recommendations for national, regional, or state initiatives. Since meeting these goals does not come cheaply, it is all the more important that energy and effort for reform also be harnessed at the local level. Schools and districts themselves, working in partnership with universities and other available expertise, need to develop examples of assessment that support their improvement efforts in elementary science education examples that will meet at least some of the criteria provided in Chapter III. At the same time, they must educate parents, school boards, and the local community to understand the severe limitations of the ubiquitous multiple-choice tests for assessing student learning and competencies in science and get these audiences to value and even demand "authentic" assessments, to borrow Archbald and Newmann's (1988) word.

As schools and districts themselves become convinced that assessments can be created that will support rather than inhibit their reform goals in elementary science education, they can carry this message to the state and national levels and demand more ecologically valid tests in science. Local development and experience with assessments of performance tasks, documentation of students' work, and records of systematic teacher observations can feed into national research and development efforts in

assessment, inform state assessments, and provide valuable material for the dissemination system we have recommended.

Throughout this report, we have proposed reforms to make both teacher-controlled and externally mandated assessments support rather than inhibit excellence in elementary science programs. Unfortunately, these reforms will not be easy to bring about. Local effort and experience is as important as national support and leadership. Success will come only if interested and committed individuals at all levels of the system take up the challenge; persist in the effort that will be needed; and share their energy, inventiveness, and expertise so as to create assessments of science learning that will adequately reflect the goals of elementary science education.

REFERENCES

Akers, J.
1984 Not all math texts are created equal. Learning 12:34-35.

Almy, Millie and Genishi, Celia.
1979 Ways of Studying Children. An Observational Manual for Early Childhood Teachers. New York: Teachers College Press.

American Association for the Advancement of Science
1985 Biology textbooks. Science Books & Films 20(5):245-286.

1986 Physics textbooks special. Science Books & Films 22(1):1-28.

1989 Science for All Americans. A Project 2061 Report. Washington, D.C.: American Association for the Advancement of Science.

Anderson, C.W.
1985 Science Testing Programs and the Improvement of Teaching. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Ill.

Anderson, C.W., and Smith, E.L.
1983 Children's Conception of Light and Color: Developing the Concept of Unseen Rays. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Anderson, J.R., Boyle, C.F., and Reiser, B.J.
1985 Intelligent tutoring systems. Science 228:450-462.

Archbald, Doug A., and Newmann, Fred M.
1988 Beyond Standardized Testing. Reston, Va.: National Association of Secondary School Principals.

Armstrong, Jane, et al.
1988 The Impact of State Policies on Improving Science Curriculum. Denver, Colo.: Education Commission of the States.

Assessment of Performance Unit
1984-1985 Science Report for Teachers: 1-6. Department of Education and Science and the Welsh Office; Department of Education for Northern Ireland. Distributed by Garden City Press Limited, Letchworth, Hertfordshire SG6 1JS, Great Britain.

Atkin, J. Myron
1980 The Government in the classroom. Daedalus 109 (Summer 1980):85-97.

Barr, Rebecca and Dreeben, Robert
1983 How School Works. Chicago: University of Chicago Press.

Baron, Joan Boykoff
1987 Evaluating Thinking Skills in the Classroom. In Joan Boykoff Baron and Robert J. Sternberg (eds.), Teaching Thinking Skills: Theory and Practice. New York: W.H. Freeman & Co.

1988 "What We Learn from State Assessments of Elementary School Science." Paper presented at the Lesley College Assessment Planning Conference, November 4-6, 1988. Available from author, Connecticut Department of Education, Hartford, Conn.

Becker, Henry Jay
1986 Instructional uses of school computers. Reports from the 1985 National Survey 1:1-12. Available from the Center for Social Organization of Schools, Johns Hopkins University.

Berman, P., and McLaughlin, M.W.
1975-1978 Federal Programs Supporting Educational Change, Vols. I-VII. R-1589/1-HEW to R-1589-8-HEW. Santa Monica, Calif.: Rand Corporation.

Blank, Rolf, and Espenshade, Pamela
1988 State Education Indicators on Science and Mathematics. Washington, D.C.: Council of Chief State School Officers.

Bloom, Benjamin S.
1984 The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher 13:4-16.

Blosser, P.E.
1984 Attitude research in science education. ERIC Clearinghouse for Science, Mathematics, and Environmental Education 1:1-18.

Blumberg, Fran, Epstein, Marion, MacDonald, Walter, and Mullis, Ina.
1986 A Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics: Final Report. Parts 1 and 2. ME-G-84-2006-P4. Princeton, N.J.: National Assessment of Educational Progress.

Bock, R. Darrell, and Mislevy, Robert J.
1988 Comprehensive educational assessment for the states: The duplex design. Educational Evaluation and Policy Analysis. 10(2):89-105.

Botkin, James W., Dimancescu, Dan, and Strata, Ray
1984 The Innovators: Rediscovering America's Creative Energy. New York: Harper and Row.

Bradburn, Norman
1979 Respondent Burden. In L. Reeder (ed.), Health Survey Research Methods: Second Biennial Conference, Williamsburg, Va. Washington, D.C.: U.S. Government Printing Office.

Bredderman, Ted
1983 Effects of activity-based elementary science on student outcomes: A quantitative synthesis. Review of Educational Research Winter, 1983, Vol. 53, No. 4:449-518.

Brown, J.S., and Burton, R.R.
1978 Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science 4:379-426.

Bybee, Rodger W., Buchwald, Edward C., Crissman, Sally, Heil, David, Matsumoto, Carolee, and McInerny, Joseph D.
1988 Science and Technology Education for the Elementary Years: Curriculum and Instruction Frameworks. Washington, D.C.: National Center for Improving Science Education.

Campione, Joseph C., and Brown, Ann L.
1987 Linking Dynamic Assessment with School Achievement. In Carol Schneider Lidz (ed.), Dynamic Assessment: An Interactional Approach to Evaluating Learning Potential. New York and London: The Guildford Press.

Carey, Neil, and Shavelson, Richard
1988 Outcomes, Achievement, Participation, and Attitudes. In Richard Shavelson, Lorraine McDonnel, and Jeannie Oakes (eds.), Indicators for Monitoring Mathematics and Science Education. A Sourcebook. Santa Monica, Calif.: The RAND Corporation.

Carrick, Tessa
1987 Reflections on practical assessment in GCSE 2: Implementation and implications of GCSE requirements. Journal of Biological Education 21(3):167-174.

Champagne, A.B., Klopfer, L.E., Solomon, C.A., and Cahn, A.D.
1980 Interactions of Students' Knowledge with Their Comprehension and Design of Science Experiments. Pittsburgh, Pa.: University of Pittsburgh, Learning Research and Development Center.

Champagne, A.B., Klopfer, L.E., and Gunstone, R.F.
1982 Cognitive research and the design of science instruction. Educational Psychologist 17(10):31-53.

Chi, M.T.H., Feltovich, P.J., and Glaser, R.
1981 Categorization and representation of physics problems by experts and novices. Cognitive Science 5:121-152.

Chittenden, Edward
1988 "Young Children's Discussion of Science Topics: Implications for Assessment and Instruction." Paper presented at the Lesley College Assessment Planning Conference, November 4-6, 1988. Available from author, Educational Testing Service, Princeton, N.J.

Clark, D.L., Lotto, L.S., and Astuto, T.A.
1984 Effective schools and school improvement: A comparative analysis of two lines of inquiry. Educational Administration Quartery 20(3):41-68.

Clune, William H., with Paula White and Janice Patterson
1989 The Implementation and Effects of High School Graduation Requirements: First Steps Toward Curricular Reform. Rutgers University, New Brunswick, N.J.: Center for Policy Research in Education.

Coffman, W.
1983 Testing in the Schools: A Historical Perspective. In E. L. Baker and J. L. Herman (eds.), Testing in the Nation's Schools: Collected Papers pp. 3-27. Los Angeles: UCLA Center for the Study of Evaluation.

Cohen, Rosalie A.
1987 A Match or Not a Match: A Study of Intermediate Science Teaching Materials. In Audrey B. Champagne and Leslie E. Hornig (eds.), The Science Curriculum. Washington, DC: American Association for the Advancment of Science.

Collins, Alan, Brown, J.S., and Newman, S.E.
in press Cognitive Apprenticeship: Teaching Students the Craft of Reading, Writing, and Mathematics. In L.B. Resnick (ed.), Cognition and Instruction: Issues and Agendas. Hillside, N.J.: Erlbaum.

Committee on Research in Mathematics, Science, and Technology Education
1985 Mathematics, Science and Technology Education: A Research Agenda. Available from the Commission on Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

1987 Interdisciplinary Research in Mathematics, Science, and Technology Education. Available from the Commission on Behavioral Social Sciences and Education. Washington, D.C.: National Academy Press.

Committee on Science, Engineering, and Public Policy
1984 High Schools and the Changing Workplace. Report of the Panel on Secondary School Education for the Changing Workplace. Washington, D.C.: National Academy Press.

Connecticut State Department of Education
1986 Connecticut Assessment of Educational Progress 1984-1985. Science Summary and Interpretations. Hartford, CT.

Council of Chief State School Officers
1984 Education Evaluation and Assessment in the United States. Position Paper and Recommendations for Action. Washington, D.C.: Council of Chief State School Officers.

Crandall, David P., and Loucks, Susan F.
1983 People, Policies, and Practices: Examining the Chain of School Improvement. Vol. X. Executive Summary: A Roadway for School Improvement. Andover, Mass.: The NETWORK, Inc.

Darling-Hammond, L., Wise, A.E., and Pease, S.R.
1983 Teacher evaluation in the organizational context: A review of the literature. Review of Educational Research 53(3):285-328.

Darling-Hammond, L. and Hudson, L.
1986 Indicators of Teacher and Teaching Quality. Santa Monica, Calif.: The RAND Corporation.

Department of Education and Science and the Welsh Office
1987 National Curriculum: Task Group on Assessment and Testing: A Report. Great Britain: Department of Education and Science and the Welsh Office.

Dorr-Bremme, Donald W., and Herman, Joan L.
1986 Assessing Student Achievement: A Profile of Classroom Practices. University of California, Los Angeles: Center for the Study of Evaluation.

Driver, R., and Oldham, V.
1986 A constructivist approach to curriculum development in science. Studies in Science Education 13:105-122.

Driver, R., Guesne, E., and Tiberghien, A.
1985 Children's Ideas in Science. Great Britain: Open University Press.

Ebel, R. L.
1967 Improving the Competence of Teachers in Educational Measurement. In J. Flynn & H. Garber (eds.), Assessing behavior: Readings in educational and psychological measurement. Reading, Mass.: Addison-Wesley.

Education Commission of the States
1982 The Information Society: Are High School Graduates Ready? Denver, Colo.: Education Commission of the States.

Ericsson, K.A., and Simon, H. A.
1984 Protocol Analysis: Verbal Reports as Data. Cambridge, Mass.: MIT Press.

Erlwanger, S.
1975 Case studies of children's conceptions of mathematics. Journal of Children's Mathematical Behavior 1(3):157-283.

Feuerstein, Reuben, Rand, Yaacov, Jensen, Morgens Reimer, Kaniel, Shlomo, and Tzuriel, David
1987 Prerequisites for Assessment of Learning Potential: The LPAD Model. In Carol Schneider Lidz (ed.), Dynamic Assessment: An Interactional Approach to Evaluating Learning Potential. New York and London: The Guilford Press.

Fleming, M., and Chambers, B.
1983 Teacher-made Tests: Windows on the Classroom. In W. Hathaway (ed.), New directions for testing and measurement: Testing in the schools pp. 29-38. San Francisco: Josey-Bass.

Frederiksen, C.H., Frederiksen, J.R., and Bracewell, R.H.
1985 Discourse Analysis of Children's Text Production. In A. Matsuhasi (ed.), Writing in Real Time. New York: Longmans.

Frederiksen, N.
1984 The real test bias: Influences of testing on teaching and learning. American Psychologist 39:193-202.

1986 Construct validity and construct similarity: Methods for use in test development and test validation. Multivariate Behavioral Research 21:3-28.

Frederiksen, N., and Ward, W. C.
1978 Measures for the study of creativity in scientific problem solving. Applied Psychological Measurement 2:1-24.

Fullilove, Robert E.
1987 "Images of Science: Factors Affecting the Choice of Science as a Career." Contractor Report. Washington, D.C.: Office of Technology Assessment.

Gardner, P.L.
1975 Attitude measurement, a critique of some recent research. Education Research 7:101-109.

Gentner, D., and Gentner, D.R.
1983 Flowing Water or Teeming Crowds: Mental Models of Electricity. In D. Gentner and A.L. Stevens (eds.), Mental Models. Hillsdale, N.J.: Erlbaum.

Glaser, Robert
1984 Education and thinking: The role of knowledge. American Psychologist 39(2):93-104.

Harvard Education Letter
1988 "Why Do Few Students Want to Become Scientists?" Harvard Education Newsletter 4(1):6.

Havelock, R.G. and Lingwood, D.A.
1973 R&D Utilization Strategies and Functions: An Analytical Comparison of Four Systems. Ann Arbor, Mich.: Institute for Social Research, University of Michigan.

Hawley, W.D., Rosenholtz, S., Goodstein, H.J., and Hasselbring, T.
1985 Good schools: What research says about improving student achievement. Peabody Journal of Education 61(4)1-178.

Hein, George E.
1987 The assessment of science learning in materials-centered science education programs. Science and Children 25(2):8-12.

Helm, H., and Novak, J.D.
1983 Proceedings of the International Seminar in Science and Mathematics. Unpublished manuscript, Ithaca, N.Y.

Hodgkinson, Harold L., ed.
1985 All One System: Demographics of Education, Kindergarten through Graduate School. Washington, D.C.: The Institute for Educational Leadership, Inc.

Horn, Elizabeth A., and Walberg, Herbert J.
1984 Achievement and interest as functions of quantity and level of instruction. Journal of Education Research 77(4):227-232.

Hudson Institute
1987 Workforce 2000: Work and Workers for the 21st Century. Indianapolis, Ind.: Hudson Institute.

Hueftle, Stacey J., Rakow, Steven J., and Welch, Wayne W.
1983 Images of Science: A Summary of Results from the 1981-82 National Assessment in Science. Minneapolis: Minnesota Research and Evaluation Center.

Human Interaction Research Institute
1976 Putting Knowledge to Use: A Distillation of the Literature Regarding Knowledge Transfer and Change. Joint project with National Institute of Mental Health. Los Angeles, Calif.: Human Interaction Research Institute.

International Association for the Evaluation of Educational Achievement
1988 Science Achievement in Seventeen Countries. A Preliminary Report. Elmsford, N.Y.: Pergamon Press.

Larkin, J., McDermott, L., Simon, D.P., and Simon, H.A.
1980 Expert and novice performance in solving physics problems. Science 208:1335-1342.

Levin, Henry, and Rumberger, Russell
1983 The Educational Implications of High Technology. Institute for Research on Education, Finance and Governance. Project Report No. 83-A4. Stanford, Calif.: Stanford University.

Little, J.W.
1982 Norms of collegiality and experimentation: Workplace conditions of school success. American Educational Research Journal 19(3): 325-340.

Lock, R., and Davies V.
1987 Assessing practical work in biology using the OCEA scheme. Journal of Biological Education 21(4):275-280.

McDermott, Lillian C.
1984 Research on conceptual understanding in mechanics. Physics Today July:240-32.

McKnight, Curtis C., Crosswhite, F. Joe, Dossey, John A., Kifer, Edward, Swafford, Jane O., Travers, Kenneth J., and Cooney, Thomas J.
1987 The Underachieving Curriculum: Assessing U.S. School Mathematics from and International Perspective. Champaign, Ill.: Stipes.

McLean, Les
1985 Drawing Implications of Instruction from Item, Topic and Classroom-Level Scores in Large-Scale Science Assessment. Paper presented at the annual meeting of American Educational Research Association, Chicago, Ill.

Messick, S., Beaton, A., and Lord, F.
1983 National Assesssment of Educational Progess Reconsidered: A New Design for a New Era. Princeton, N.J.: Educational Testing Service.

Mullis, Ina V.S., and Jenkins, Lynn B.
1988 The Science Report Card. Elements of Risk and Recovery. National Assessment for Educational Progress. Report No. 17-S-01. Princeton, N.J.: Educational Testing Service.

Munby, Hugh
1983 Thirty studies involving the Scientific Attitude Inventory: What confidence can we have in this instrument? Journal of Research in Science Teaching 20(2): 141-162.

Murnane, Richard J., and Raizen, Senta A., eds.
1988 Improving Indicators of the Quality of Science and Mathematics Education in Grades K-12. Committee on Indicators of Precollege Science and Mathematics Education, National Research Council. Washington, D.C.: National Academy Press.

NAEP (National Assessment of Educational Progress)
1987 Learning by Doing. Report No.: 17-HOS-80. Princeton, N.J.: Educational Testing Service.

National Alliance of Business
1987 The Fourth R: Workforce Readiness. Washington, D.C.: National Alliance of Business

National Commission on Excellence in Education
1983 A Nation at Risk: The Imperative for Educational Reform. Supt. of Doc. No. 065-000-00177-2. Available from the U.S. Government Printing Office. Washington, D.C.: U.S. Department of Education.

National Governors' Association
1987a Making America Work. Washington, D.C.: National Governors' Association.

1987b The Role of Science and Technology in Economic Competitiveness. Washington, D.C.: National Governors' Association.

National Science Board
1987 Science and Engineering Indicators - 1987. NSB 87-1. Washington, D.C.: National Science Foundation.

National Science Board Commission on Precollege Education in Mathematics, Science and Technology
1983 Educating Americans for the 21st Century. CPCE-NSF-03. Washington, D.C.: National Science Foundation.

National Science Resources Center
1986 National Conference on the Teaching of Science in Elementary Schools. Summary. Available from the National Science Resources Center, National Academy of Sciences-Smithsonian Institution, Washington, D.C.

1988 Science for Children. Washington, D.C.: National Academy Press.

Nickerson, Raymond S.
1988 On improving thinking through instruction. Review of Research in Education 15:3-57.

Nuthall, G., and Lee, A.A.
1982 Measuring and Understanding the Way Children Learn in Class. Technical Report: Teaching Research Project. Education Department, University of Canterbury, Christchurch, New Zealand.

Oakes, Jeannie
1986 Educational Indicators: A Guide for Policymakers. OPE-01. Santa Monica, Cal.: Rand Corporation.

in press What Educational Indicators? The Case for Assessing the School Context. Educational Evaluation and Policy Analysis.

Office of Technology Assessment
1988 Power On! New Tools for Teaching and Learning. Available from U.S. Government Printing Office. Report No. OTA-SET-379. Washington, D.C.: Congress of the United States, Office of Technology Assessment.

Osborne, R.J., and Wittrock, M.C.
1983 Learning science: A generative process: Science Education 67:489-508.

Penick, John E., ed.
1983 Focus on Excellence: Elementary Science. Vol. 1, No. 2. Washington, D.C.: National Science Teachers Association.

Pine, Jerome
1988 "Evaluation of Alternatives for Assessing Science Process Learning by Elementary School Children." National Science Foundation grant SPA-8751511. Further information available from author, California Institute of Technology.

Purkey, Stewart C., and Smith, Marshall S.
1983 Effective schools -- a review. Elementary School Journal 83(4):426-452.

Raizen, Senta A.
1987 Assessing the Quality of the Science Curriculum. In Audrey B. Champagne and Leslie E. Hornig (eds.), The Science Curriculum. Washington, D.C.: American Association for the Advancement of Science.

Raizen, Senta A., and Jones, Lyle V., eds.
1985 Indicators of Precollege Education in Science and Mathematics. A Preliminary Review. Committee on Indicators of Precollege Science and Mathematics Education, National Research Council. Washington, D.C.: National Academy Press.

Resnick, Lauren B.
1983 Mathematics and science learning: A new conception. Science 220(4):477-478.

1987 Education and Learning to Think. Washington, D.C.: National Academy Press.

Rogers, E.M.
1962 Diffusion of Innovations. New York: Free Press.

Rosenholtz, S.J.
1985 Effective schools: Interpreting the evidence. American Journal of Education 93:352-388.

Rowe, Mary Budd
1979 Externality and Children's Problem Solving Strategies. Paper prepared under NIMH grant no. R01MH 25229. Available from the University of Florida, Gainesville.

Rudman, H., Kelly, J.L., Wanous, D.S., Mehrens, W.A., Clark, C.M., and Porter, A.C.
1980 Integrating Assessment with Instruction: A Review 1922-1980. East Lansing, Mich: Institute for Research on Teaching.

Rutter, M.
1983 Effective Schools. In Lee Shulman and Gary Skyes (eds.), Handbook of Teaching and Policy. New York: Longman.

Scheuer, J.H.
1987 Competitiveness and the Quality of the American Workforce. Opening Statement, Hearings before the Subcommittee on Education and Health of the Joint Economic Committee, United States Congress, October 1, 1987, Washington, D.C.

Shamos, Morris
1988 The lesson every child need not learn. The Sciences July/Aug 1988:14-20.

Sharp, Laure M., and Frankel, Joanne
1983 Respondent burden: A test of some common assumptions. Public Opinion Quarterly 47:36-53.

Shavelson, Richard, McDonnel, Lorraine, Oakes, Jeannie, and Carey, Neil
1987 Indicator Systems for Monitoring Mathematics and Science Education. Report No. R-3570-NSF. Santa Monica, Calif.: The RAND Corporation.

Shymansky, James A., Kyle, William C., Jr., and Alport, Jennifer M.
1983 The effects of new science curricula on student performance. Journal of Research in Science Teaching 20(5):387-404.

Sinaiko, H.W., and Broedling, L.A.
1976 Perspectives on Attitude Assessment: Surveys and Their Alternatives. Champaign, Ill.: Pendelton Publications.

Sizer, Theodore
1984 Horace's Compromise: The Dilemma of the American High School. Boston: Houghton Mifflin.

Snow, Richard E.
1980 Aptitude and Achievement. In W.B. Schrader (ed.), Measuring Achievement: Progress over a Decade. New Directions for Testing and Measurement. San Francisco: Jossey-Bass.

Stake, Robert E., and Easley, Jack A. Jr.
1978 Case Studies in Science Education. NSF SE-78-74. Available from the U.S. Government Printing Office. Washington, D.C.: National Science Foundation.

Stevens, A., Collins, A., and Goldin, S. E.
1979 Misconceptions in students' understanding. International Journal of Man-Machine Studies 11:145-156.

Stevenson, Harold W., Lee, Shin-Ying, and Stigler, James W.
1986 Mathematics achievement of Chinese, Japanese, and American Children. Science 231:693-699.

Strike, K.A.
1982 Educational Policy and the Just Society. Urbana, Ill.: University of Illinois Press.

Task Force on Education for Economic Growth
1983 Action for Excellence: A Comprehensive Plan to Improve Our Nation's Schools. Denver, Colo.: Education Commission of States.

Twentieth Century Fund
1983 Report of the Twentieth Century Fund Task Force on Federal Elementary and Secondary Education Policy. New York: The Twentieth Century Fund.

U.S. Department of Education
1988 Creating Responsible and Responsive Accountability Systems: Report of the OERI State Accountability Study Group. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Practice.

Viginia Department of Education
1986 Science Education Program Assessment Model Resource Guide. Creating a More Effective Learning Environment. Richmond, Va.: Virginia Department of Education.

Ward, W.C., Frederiksen, N., and Carlson, S.
1980 Construct validity of free-response and multiple-choice versions of a test. Journal of Educational Measurement 17:11-29.

Watts, D.M., and Gilbert, J.K.
1983 Enigmas in school science: Students' conceptions for scientifically associated words. Research in Science and Technology Education 1(2):161-171.

Weiss, Iris S.
1978 Report of the 1977 National Survey of Science, Mathematics and Social Studies Education. Prepared for the National Science Foundation. Supt. of Doc. No. 083-000-00364-0. Available from the U.S. Government Printing Office. Washington, D.C.: National Science Foundation.

1987 Report of the 1985-86 National Survey of Science and Mathematics Education. Prepared for The National Science Foundation. No. SPE-8317070. Available from the U.S. Government Printing Office. Washington, D.C.: National Science Foundation.

Welch, Wayne
1984 A Science-based Approach to Science Learning. In David Holdzkom and Pamela Lutz (eds.), Research Within Reach: Science Education. Charleston, W.Va.: Appalachia Regional Laboratory.

White, Barbara Y., and Horwitz, Paul
1987 TinkerTools: Enabling Children to Understand Physical Laws. BBN Report No. 6470. Cambridge, Mass.: BBN Laboratories, Inc.

Willson, V.L.
1983 A meta-analysis of the relationship between science achievement and science attitudes: Kindergarten through college. Journal of Research in Science Teaching 20(9):839-850.

Woellner, R. S.
1979 Let's use tests for teaching: Standardized test results can provide the basis for a program of instruction. Teacher 90(2):62-64,179-181.

Woolnough, B.E., & Allsop, T.A.
1985 Practical Work in Science. Cambridge, England: Cambridge University Press.

Womer, F.B.
1981 State-level Testing: Where We Have Been May Not Tell Us Where We Are Going. In D. Carlson (ed.), New Directions for Testing and Measurement: Testing in the States: Beyond Accountability. San Francisco: Jossey-Bass.

Yeh, J. P., Herman, J. L., and Rudner, L. M.
1981 Teachers and Testing: A Survey of Test Use. CSE Report No. 166. University of California, Los Angeles: Center for the Study of Evaluation.

Yin, R.K., Heald, K.A., Vogel, M.E., Fleischauer, P.D., and Vladeck, B.C.
1976 A Review of Case Studies of Technological Innovations in State and Local ervices. R-1870-NSF. Santa Monica, Calif.: Rand Corporation.

APPENDIX

APPENDIX

(From the Center's Report: Science and Technology Education for the Elementary Years: Frameworks for Curriculum and Instruction)

FUNDAMENTAL ORGANIZING CONCEPTS FOR ELEMENTARY SCHOOL SCIENCE

A paradox arises when schools try to prepare students for the future. Most educators are convinced of two equally valid but contradictory ideas; the world is changing at an accelerating pace, and there are fundamental, enduring concepts for organizing thinking about the world.

The trite saying, "the only constant in modern life is change" is a poor description because change itself is occurring at faster rates and in different directions. If this is true, then what should be taught in elementary school science that will have lasting value to students? What will help them understand and adjust to change? Are there explanatory concepts that are so fundamental and powerful that they will always be valid and useful? We think there are some. There are fundamental organizing concepts in science that all students, by the time they finish sixth grade, should incorporate in the way they think about and engage their world. These concepts are valuable because

- they are applicable to both science and technology,
- they have applications beyond science and technology,
- they accommodate different developmental levels,
- they apply to the personal lives of children, and,
- they are powerful explanatory concepts.

Organization (or orderliness)

Ideas and descriptions about the world can be organized in different ways including hierarchies, simple-to-complex arrays, and symmetry. Objects in nature or the classroom can be assembled into groups showing hierarchies, such as atoms, molecules, mineral grains, rocks, strata, hills, mountains, and planets. Some organisms contain hierarchies in themselves like the trunk, branches, twigs, stems, and leaves of trees or the hierarchies within social systems, such as transportation or communication.

Varieties of organisms from single-celled amoeba, to sponges, to corals, and so on, can illustrate simple-to-complex arrays. Technology provides examples of increasingly complex objects that serve similar purposes. As an illustration, people slide down hills in the winter using sheets of plastic, or they use toboggans, sleds, or aerodynamic bobsleds. The objects are increasingly sophisticated, but all are designed to carry passengers on a thrilling downhill ride.

Objects can be described according to common elements of symmetry and polarity: they possess top and bottom, front and back; and in many cases, shapes are repeated when the objects or organisms are turned or inverted.

TABLE 1

TEACHING EXAMPLES
FOR **ORGANIZATION**

PRIMARY (K-3) INTERMEDIATE (4-6)	
• Sorting objects (e.g., objects that sink and objects that float)	• Identifying levels of organization, such as atoms; molecules; cell-tissue-organs; earth-solar system; stars-galaxies; and organism, population, community, ecosystem
• Ordering events (e.g., identifying the order of planting a seed, sprouting, adult plant, flower, and fruit)	• Describing the component parts of natural and technological systems
• Classifying objects and organisms	• Specifying the hierarchial relationship among parts of natural and technological systems
• Identifying groups of similar animals (e.g., mammals, reptiles, insects)	• Describing the constituents of rocks
• Identifying groups of similar plants (e.g., beans, grass, roses)	• Recognizing patterns of leaves
• Developing a simple scheme for classifying objects or organisms (e.g. animals typically found in certain environments)	• Identifying geometric shapes
• Classifying objects and organisms from simple to complex	• Describing symmetry of objects and organisms
• Identifying solids, liquids, and gases (e.g. water as ice, water, and vapor)	• Dismantling and reassembling a simple machine
• Identifying groups of objects that have been designed or constructed by humans	• Recognizing organization within and among the atmosphere, hydrosphere, lithosphere, and celestial sphere

Cause and Effect

Nature behaves in ways that are predictable. Searching for causes and explanations is the major activity of science; effects cannot happen without causes. A common error arises when individuals assume that events that occur simultaneously or sequentially have a cause-and-effect relationship. For example, the rotation of the planets and a death in one's family, or a pregnant woman's sighting of a rabbit and the birth of a child with cleft lip may happen simultaneously, but there is not causal interrelationship. Some events require multiple causes, that is, several things must happen to cause an effect.

Classic activities with seed growing can illustrate cause and effect concepts. For beans to be healthy, seeds need water, light, and warmth; well-organized experiments can show the effect of varying each of these three parameters.

Cub Scouts discover that streamlining, carefully aligned axles, and good lubrication all help to make a pinewood derby car run faster. They also discover that if too much wood is carved off the car body when attempting to make it streamlined, weight must be added to keep it heavy. There are optimum conditions for optimum performance.

TABLE 2

TEACHING EXAMPLES
FOR **CAUSE AND EFFECT**

PRIMARY (K-3) INTERMEDIATE (4-6)	
• Describing health risks (e.g., riding a bicycle, crossing streets)	• Identifying the effects of poor nutrition
• Identifying changes, e.g., heating/cooling, moving/not moving	• Describing cause and effect in simple activities such as growing seeds
• Describing simple technologies (e.g., scissors, paper clips, pencils)	• Describing the effects of various substances on objects and organisms
• Using everyday examples to describe cause and effect (e.g., lights, water, temperature)	• Designing simple machines that achieve a desired effect
• Predicting a sequence of events for natural phenomena and technological objects	• Describing natural phenomena in terms of cause and effect (e.g., weather, erosion)
• Describing interactions between objects and organisms (e.g., eating is related to growth and development)	• Differentiating between correlation and cause and effect
	• Giving evidence for interactions between and among simple systems

Systems

Systems consist of matter, energy, and information that move about from reservoir to reservoir through carefully delimited pathways. The amount of matter, energy, information in reservoirs, and the rate of transfer through pathways varies over time. Systems are understood by tracking changes and drawing boundaries around the constituent parts.

One of the best known natural systems is the hydrologic cycle. Water in solid, liquid, and gaseous phases moves about the earth's surface sometimes residing in the atmosphere, sometimes in living tissue, and sometimes in streams, lakes, groundwater, and oceans. Being able to observe and measure this system helps us understand weather, water supply, and pollution.

In the classroom, an aquarium might serve as a system. To make it a balanced aquarium, the plants have to use fish waste products to provide enough oxygen and food for the fish to survive. Of course, the plants also depend on a light source for photosynthesis. Balancing the aquarium requires some knowledge about the matter and energy present and how it follows pathways from plants to water to animals.

Most technology can be seen as systems. A common example is the furnace and thermostat. This system is cybernetic; that is, information is related and acted upon within the system in a stabilizing way. A properly tuned heating system keeps room temperatures from fluctuating more than a few degrees from the set point.

TABLE 3

TEACHING EXAMPLES
FOR SYSTEMS

PRIMARY (K-3) INTERMEDIATE (4-6)	
• Describing whole systems, such as toys and simple machines	• Taking apart simple machines
• Exploring a simple natural system	• Describing the school's transportation system
• Constructing a simple technological device	• Differentiating systems and subsystems
	• Applying the concept of systems to different objects, events, and organisms (e.g., humans, earth, electrical)
	• Describing the characteristics of different natural and technological systems, (i.e., the boundaries, components, feedback, resources)
	• Identifying matter and energy as essential to systems

Scale

Scale refers to quantity in both a relative and an absolute sense. Thermometers, rulers, and weighing devices help students to see precisely that matter and energy vary in quantity. Notions of scale in an absolute sense are important because in the physical and biological world certain phenomena happen only within fixed limits of size.

For instance, in biology, water striders are superbly scaled; they are able to run across a puddle suspended by the surface tension of water. If water striders were much larger, they would sink; if they were much smaller and became wet, they would not be able to break away from the clinging water. Full-term newborn babies are not healthy if they are very large or very small. There is an ideal size range for healthy babies.

In technology, scale is important to efficient operation. Buses may only get 5 or 6 miles per gallon, but they can carry 40 or 50 passengers, thus making them far more fuel efficient than passenger cars. However, technological devices must also account for human scale. The bus driver's seat must be designed to accommodate tall, medium, and short drivers.

TABLE 4

TEACHING EXAMPLES
FOR SCALE

PRIMARY (K-3)
INTERMEDIATE (4-6)

- Drawing simple objects in actual size and comparing the drawing to scale pictures
- Recognizing the differences in children and adults
- Knowing that some objects, such as doll houses and toy trucks are scale models of real objects
- Designing a model of a simple object or organism
- Defining big/little, near/far, short/long
- Stating different scales of time, space, and matter
- Mapping a small area
- Describing the magnification on a microscope in terms of scale
- Making a solar system to scale for both size of planets and distance
- Estimating the size of an object
- Computing the scale of geologic time and astronomic distance
- Designing a machine and then building the machine

Models

To make sense of the world around them, human beings create models or metaphors that show the essential character of the phenomena that interest them. Furthermore, the models may be conceptual (consisting of word descriptions or drawings), mathematical (consisting of equations or other formal representations); or physical (consisting of real objects that possess some of the characteristics of the real thing).

The solar system is often modeled in the classroom by describing the planets as huge balls moving about an even larger sun. Such a model solar system is usually to scale for both size of planets and distance between planets. A mathematical model of the solar system might include the shape of a planet's orbit as being elliptical. And finally, a physical model of the solar system might consist of a series of scale-sized balls placed at appropriate distances throughout the room or hallway.

Models often serve as prototypes in technology and in that case may be full-sized representations of the final product. Models usually possess only some of the characteristics of the real thing. Children readily understand that most toys are models that look like real objects, such as cars, airplanes, babies, and animals, but do not possess all the attributes of those objects.
Models can be used to test the workings of technology without costly investments in full-scale objects. Small boats and airplanes are tested in tanks and wind tunnels before their full-sized counterparts are built. In this way, many design experiments can be tested inexpensively to find optimum results.

TABLE 5

TEACHING EXAMPLES
FOR MODELS

PRIMARY (K-3)
INTERMEDIATE (4-6)

- Recognizing numbers as representations of objects or organisms
- Describing the differences between a toy car and a real car
- Providing a picture of a car or person
- Identifying models that are bigger than, smaller than, or the same size as the real object or organism and explaining why each is useful
- Constructing a simple graph
- Representing graphically a relationship such as color and wavelength
- Differentiating between a model and reality
- Constructing models of linear and exponential growth

Change

Change is continuing and ubiquitous in the natural world. Some objects or organisms (species) seem unchanging, but that is a function of humans' inability to perceive the rate or scale of change. For example, mountains erode and species evolve, but the time required to recognize substantial change is quite long. Changes in the size and structure of the universe are too large for human beings to observe and to measure directly, and mutations in genetic material are hidden unless they affect observable characteristics.

Change in the natural world generally tends toward disorganization unless energy is put back into the system. For example, a child's well-organized bedroom will tend toward clutter (a mess) unless energy is expended to keep the room organized. Similarly, a bicycle will tend toward disrepair and wear out unless energy is expended to maintain it.

Some change is cyclical; that is, the direction of the change is reversed. Diurnal cycles, lunar cycles, seasonal cycles, and menstrual cycles are examples. Some change is one-directional; physical growth and intellectual development, puberty, and menopause for example.

The rate of change can vary. For example, although all (normal) sixth graders will ultimately progress through the same developmental stages, not all of them will reach the same developmental landmarks at the same time.

Technology changes as new problems arise and as new solutions supplant old. Historically, many technologies have become more complex and have changed from functional adaptation to convenient utilization.

TABLE 6

TEACHING EXAMPLES
FOR **CHANGE**

PRIMARY (K-3) INTERMEDIATE (4-6)	
• Identifying the different seasons by their attributes	• Naming the stages of development
• Observing and describing immediate changes	• Observing and describing the properties of water, as in solid to liquid to gas
• Observing delayed changes	• Observing and recording the phases of the moon
• Observing personal changes in a day, week, year	• Identifying the changes in an ecosystem
• Identifying different types and rates of change	• Investigating different life cycles
• Describing growth of organisms	• Estimating the rate and direction of simple changes in physical systems
• Identifying indications of seasonal change during a nature walk	• Differentiating between linear and exponential growth
	• Recognizing the limits of change in simple systems

Structure and function

There is a relationship between the way organisms and objects look (feel, smell, sound, taste) and the actions they perform. The structure of leaves, for example, is related to their functions of energy production and transpiration. Scent glands in skunks are related to protection. All automobiles have a similar shape because engineers know that this shape improves the ability of an automobile to move down the highway efficiently. Similarly, round, inflatable tires on a bicycle are conducive to the bicycle's function. More specifically, light-weight tires are designed for racing and knobby tires are better for all-terrain bikes where traction is important.

In the biological world, both structure and function are results of cumulative natural selection, the major mechanism of organic evolution. The relationship is not a function of purposeful design, nor does it occur by accident (unless one considers the accidental nature of mutation, which is the ultimate source of all variations that may have adaptive function).

The structure/function relationship also appears in artifacts. Archaeologists explain artifacts by determining the functions of various shapes and forms found. For example, small arrowheads were used for hunting birds, large spear heads were used for larger animals. Some stones look and feel like scrapers or hammers and most certainly must have been used for those purposes. The congruence between structure and function in the designed world (technology) is purposeful. Furthermore, the congruence can be refined by experimentation.

TABLE 7

TEACHING EXAMPLES
FOR STRUCTURE AND FUNCTION

PRIMARY (K-3)
INTERMEDIATE (4-6)

- Observing the structure of an animal and its relationship to function
- Describing the function of a simple system (e.g., roof shape for shedding rain and snow)
- Designing a common object, such as a plate, bowl, spoon, or fork
- Examining simple plants and describing the parts and functions
- Describing a bicycle in terms of structure and function
- Building a structure from simple materials
- Designing a plant or animal
- Inventing a simple device for measuring wind velocity
- Interpreting antique objects
- Interpreting animal tracks
- Recognizing the relationship of structure and function in humans, buildings, environments
- Describing the functions of human body parts
- Describing the structure and function of tools
- Recognizing the abiotic and biotic structures of an ecosystem

Discontinuous and Continuous Properties (Variations)

All organisms and objects have distinctive properties. Variation is a universal characteristic of the natural world. Some properties are so distinctive that no continuum connects them. Examples of such discontinuous properties are living/nonliving and saltiness/sweetness.

Most properties in the natural world vary continuously; that is, there is no clear demarcation that distinguishes the variation in a population or the properties of objects. The colors of the spectrum, for example, constitute a continuum. Night and day, height, weight, resistance to infection, and intelligence are all continuous properties.

Discontinuous variation lends itself to classification of objects by type; this kind of classification emphasizes general properties rather than specific characters. Continuous variation, on the other hand, makes typological classification difficult, because it (continuous variation) emphasizes finely graded, individual distinction, as well as unity of pattern. An understanding of continuous variation is the basis of thinking about populations and is essential to an understanding of organic evolution and the statistical nature of the world.

TABLE 8

TEACHING EXAMPLES
FOR **DISCONTINUOUS AND CONTINUOUS PROPERTIES**
(VARIATIONS)

PRIMARY (K-3) INTERMEDIATE (4-6)	
• Observing different tones of colors (e.g., variations of blue)	• Investigating the changes and continuity in properties in a life cycle
• Listening to different sounds	• Recognizing the continuous properties of color in a spectrum
• Differentiating living and non-living	• Analyzing a graph of height in class--contrast with histogram of boys and girls
• Exploring the properties of objects that sink and float	• Sampling height of individuals over time
• Developing a growth chart over time	• Differentiating between day and night
	• Describing the on-off switch as a discontinuous variable

Diversity

Diversity is perhaps the most obvious characteristic of the natural world. Not only are there many different types of objects and organisms but there also is considerable variation within those objects and organisms.

As scientific understanding of the natural world has improved, humans have come to see that maintenance of diversity is important to natural systems. For example, trees, rocks, and people all play important parts in the ecological balance of a tropical rain forest. Should one component be eliminated, the entire rain forest is likely to suffer.

Technology proposes diverse solutions to problems of human adaptation to the environment. Snowshoes, cross country skis, and snowmobiles are diverse solutions to the problem of moving people across the snow. Such issues as economics, efficiency, and esthetics will help determine which solution is best.

Diversity also is evident in human values and ideas. This diversity influences the problems individuals and societies choose to address.

TABLE 9

TEACHING EXAMPLES
FOR **DIVERSITY**

PRIMARY (K-3)
INTERMEDIATE (4-6)

- Observing objects and developing a simple classification scheme
- Observing different types of objects and organisms
- Identifying the differences in pets
- Observing and describing the differences among students in class
- Listing the natural objects and organisms on the school grounds
- Listing the constructed objects on the school grounds
- Collecting organisms or objects
- Observing the differences among leaves
- Analyzing height and weight distribution among class members
- Identifying the range of similar rocks, animals, or plants
- Studying a simple ecosystem to identify the diversity of organisms
- Describing the components of similar physical systems such as airline and automobile travel
- Observing the variations within one type of leaf
- Developing a life list of birds
- Making a collection of minerals and rocks

The National Center for Improving Science Education, funded by the U.S. Department of Education's Office of Educational Research and Improvement, is a partnership of The NETWORK, Inc. and the Biological Sciences Curriculum Study (BSCS). Its mission is to promote changes in state and local policies and practices in science curriculum, science teaching, and the assessment of student learning in science. To do so, the Center synthesizes and translates recent and forthcoming studies and reports in order to develop practical resources for policymakers and practitioners. Bridging the gap between research, practice, and policy, the Center's work promotes cooperation and collaboration among organizations, institutions, and individuals committed to the improvement of science education.

The NETWORK, Inc.

BSCS
The Natural Selection